



Statistical Analysis of Indicators of Change in Pesticide Residues (Pilot Study)

Research Project conducted on behalf of
the Food Standards Agency

Final Report (March 2007)

Andrew Mead
Biometrician (Warwick HRI)

This report is dedicated to the memory of Dr. Neil Kift (NFU, and previously Horticulture Research International), who was instrumental in the initial development of the project, but tragically died in a car accident in the summer of 2006 before the project was completed.

Contents

Summary of Project	1
Background	1
Aims and Objectives	1
Data	1
Methodology	2
Results	3
Secondary data	3
Power analysis	4
Extensions	5
Conclusions	6
Detailed Description of the Project	7
Background	7
Data	8
<i>Table 1: Apple pesticide residue data sets</i>	8
<i>Table 2: Strawberry pesticide residue data sets</i>	8
<i>Table 3: Protected lettuce pesticide residue data sets</i>	9
Methodology	9
Results	11
Initial Data Summary	11
Apples	11
Strawberries	11
Protected lettuce	12
Results from New Approach	12
Apples	12
<i>Table 4: Carbendazim residues in apples</i>	13
<i>Figure 1: Distribution of carbendazim residues in apples...</i>	14
<i>Figure 2: Distribution of carbendazim residues in UK apples...</i>	14
<i>Figure 3: Distribution of carbendazim residues in imported apples...</i>	15
<i>Figure 4: Distribution of carbendazim residues in apples...</i>	16
<i>Table 5: Chlorpyrifos residues in apples</i>	17
<i>Figure 5: Distribution of chlorpyrifos residues in apples...</i>	17
Strawberries	18
Protected lettuce	18
<i>Table 6: Iprodione residues in protected lettuce</i>	19
<i>Figure 6: Distribution of iprodione residues in protected lettuce...</i>	19
<i>Figure 7: Distribution of iprodione residues in protected lettuce...</i>	20
Comparisons with Existing Approaches	21
<i>Table 7: Quintozene residues in protected lettuce</i>	22
<i>Figure 8: Distribution of quintozene residues in protected lettuce...</i>	23
Limitations of the approach	23
Secondary Data	24
<i>Table 8: Apples – pesticide usage data for pesticides applied in orchards...</i>	24
<i>Table 9: Apples – pesticide usage data for pesticides applied post harvest...</i>	24

Table 10: Strawberries - pesticide usage data...	25
Power Analysis	25
Approach	25
Distributions, Thresholds and Samples	26
Figure 9: Expected proportions of observations... for six different Gamma distributions (different mean values with variance = mean)	26
Figure 10: Expected proportions of observations ... for six different Gamma distributions (different mean values with variance = 0.1)	27
Figure 11: Expected proportions of observations ... for six different Gamma distributions (different mean values with variance = 0.8)	28
Results	28
Comparison of samples drawn from the same distribution	28
Table 11: Simulation of two samples drawn from Gamma distributions with equal means, and mean = variance = 0.1	29
Table 12: Example samples of size 60 drawn from a Gamma distribution with mean = variance = 0.1	29
Table 13: Simulation of two samples drawn from Gamma distributions with equal means, and mean = variance = 0.4	30
Table 14: Simulation of two samples from Gamma distributions with equal means, and mean = variance = 0.8	30
Table 15: Simulation of two samples drawn ...with equal means, mean = variance and sample size = 40	31
Table 16: Simulation of three samples drawn ...with equal means, mean = variance and sample size = 40	31
Table 17: Simulation of five samples drawn ...with equal means, mean = variance and sample size = 40	32
Comparisons of samples drawn from distributions with different means but the same variances	32
Figure 12: Simulated power curves for the comparisons of two samples of size 20 both ... with variance = 0.1	33
Figure 13: Simulated power curves for the comparisons of two samples of size 80 both ... with variance = 0.1	34
Figure 14: Simulated power curves for the comparisons of two samples of the same size, both ... with variance = 0.1	35
Figure 15: Simulated power curves for the comparisons of two samples of size 20, both ... variance = 0.4	36
Figure 16: Simulated power curves for the comparisons of two samples of the same size, both ... with variance = 0.4	37
Comparison of multiple samples drawn from distributions with different means but the same variance	37
Table 18: Simulation of three samples drawn from Gamma distributions with variance = 0.1 and sample size = 20	38
Table 19: Simulation of three samples drawn from Gamma distributions with variance = 0.4 and sample size = 20	38
Table 20: Simulation of three samples drawn from Gamma distributions with variance = 0.4 and sample size = 60	39
References	39

Summary of Project

Background

Previously the only quantitative analysis of trends in pesticide residues has been to consider the proportions of residues below the limit of determination (LOD), and even here, little formal statistical analysis has been applied. Recent research on the impact of insecticide resistance on pest management in lettuce resulted in the development of a novel statistical approach to the non-parametric comparison of the distributions of the numbers of aphids per plant between different treatments (Kift, Mead *et al*, 2004). This novel approach considers the proportions of observations in a number of non-overlapping, contiguous intervals, and overcomes problems of overdispersion associated with the conventional log-linear analysis of the total count per plant or plot. Where the conventional analysis failed to detect treatment effects, this novel approach allowed obvious treatment effects to be detected as statistically significant. Discussions between Neil Kift (NFU, ex-Horticulture Research International) and the Food Standards Agency suggested that this novel statistical approach might be applied to the analysis of pesticide residue data, by combining information about the distribution of residues above the limit of determination with that about the number of residues below the limit of determination.

Having helped to initiate this project, Neil Kift tragically died in a car accident in the summer of 2006, before the project was completed, and was therefore unable to contribute to the project, or comment on this report.

Aims and Objectives

The aims and objectives of this pilot study are:

- To assess whether this novel statistical approach can detect differences in pesticide residue distributions using test data sets for apples, strawberries and protected lettuce;
- To compare the results of this new analysis approach with results for simpler or more conventional analysis methods;
- To determine whether detected trends in pesticide residue distributions can be related to information about pesticide usage and pest pressure;
- To assess the power of the novel statistical approach to detect changes in the distribution of pesticide residues, and the influence on the power of the choice of classifying thresholds, sample size, and the proportion of residues below the limit of detection.

Data

Three sets of pesticide residue data were provided by the Food Standards Agency to be used as test datasets in this study:

- Apples: data from seven years (1998 – 2004) with residue levels measured for up to 6 pesticides (bupirimate, captan, carbendazim, chlorpyrifos, diphenylamine, metalaxyl), with the source of the assessed crop being identified as either UK or Imported (divided into EC and non-EC in 2004) or Unknown;
- Strawberries: data from four years (1999, 2000, 2001, 2004) with residue levels measured for up to 4 pesticides (bupirimate, iprodione, pyrimethanil, fenhexamid), with the source of the crop being identified as either UK or Imported (divided into EC and non-EC in 2004);

- Protected lettuce: data from 10 sampling occasions (over five years from 2002 – 2006, samples in both spring and autumn) with residue levels measured for up to 9 pesticides (iprodione, propyzamide, azoxystrobin, dithiocarbamates, cypermethrin, inorganic bromide, quintozone, pyrimethanil, fenhexamid).

Methodology

An initial data summary calculated the proportion of non-zero residues (i.e. above the limit of determination), the mean non-zero residue level, and the maximum recorded residue level for each pesticide for each sampling occasion.

Formal analyses were performed for datasets constructed for each different pesticide within a crop.

Two simple/conventional analyses were considered to provide base-line results against which the result from the new method could be compared:

- Analysis of variance (ANOVA) of all residue data – whilst the assumptions of normality and homogeneity of variance almost certainly fail for every dataset, this still provides a useful (and simple) base-line result;
- Analysis of the proportions of residues above and below the limit of determination (LOD) by analysing a contingency table of the numbers of residues classified by treatment (sampling occasion/source) and “above/below LOD”, using a log-linear model (a generalised linear model (GLM) assuming a Poisson error distribution and logarithm link function).

This latter approach is essentially the simplest version of the novel statistical approach, with just one “threshold” value, the value of the LOD, used to categorise the residues. Where the LOD varied between sampling occasions, the observed residues were re-evaluated relative to the highest LOD so that the proportional response relative to a consistent LOD was obtained.

The novel analysis method also analyses a contingency table of the numbers of residues using a log-linear model, but with the residues classified according to a set of thresholds. The choice of this threshold set is important, and could impact on the results of the analysis. Each threshold set considered included the LOD as the lowest threshold, with other thresholds chosen as multiples of the LOD following a logarithmic-type scale. For two data sets with a large proportion of the residues above the LOD, the effect of using an alternative set of thresholds based on the quartiles of the observed residues (as used in Kift, Mead *et al*, 2004) was also assessed.

A formal assessment of the additional discrimination provided by the multiple threshold analysis method compared with the simpler “above/below LOD” approach was possible by considering the additional variability explained by the more complex threshold set.

Two graphical approaches were considered to aid interpretation of the results of the analysis. Cumulative relative frequency plots show the different shapes of the approximate distributions for different sampling occasions (as developed in Kift, Mead *et al*, 2004). Stacked bar-charts display the relative proportions of residues in each classification group, and also provide a visual method for comparing the simpler and more complex threshold sets.

Results

An initial summary of the data suggested that the novel analysis method might be able to detect differences in the distributions of residues for a number of pesticides for each of the three crops. A potentially limiting factor was the proportion of recorded residues below the limit of determination (LOD) for some pesticides – where the vast majority of residues are below the LOD, it is unlikely that any analysis approach will be able to detect changes. Another potential problem was the variability in LOD values between samples – in most cases datasets were “re-evaluated” to the highest LOD used for a particular pesticide, so that a consistent LOD could be used. In some cases, one or more of the sampling occasions was omitted from the analysis because of a much higher LOD, use of which would remove most of the information about the distribution of residues.

For apples, four of the pesticide datasets (for captan, carbendazim, chlorpyrifos, and diphenylamine) were analysed, and all analyses indicated either effects of year or source (UK, Imported) or both, and some analyses indicated interaction effects.

For strawberries, all four of the pesticide datasets were analysed, and, again, analyses indicated either effects of year or source or both, and, for Iprodione, an interaction effect.

For protected lettuce, all nine of the pesticide datasets were analysed, but the analyses only indicated significant effects of occasion for four of the pesticides (iprodione, propyzamide, cypermethrin, quintozone).

For some datasets, alternative classification thresholds were used to define the contingency tables to be analysed, and the significance of effects did vary depending on the chosen set of thresholds.

The cumulative relative frequency curves did not prove particularly useful for identifying the causes of any significant effects, possibly because many of the differences in distributions between sampling occasions were primarily associated with the percentage of residues below the LOD. However, the alternative graphical approach, using stacked bar-charts, does appear to provide a useful graphical summary of the differences between distributions, as well as indicating where the use of a more complex set of multiple thresholds might be of additional benefit compared with the simpler “Below/Above LOD” threshold analysis.

The formal comparison of analyses using the more complex set of multiple thresholds with those using the single LOD threshold showed that there was evidence for all four apple pesticide data sets analysed, for the strawberries iprodione data set, and for the protected lettuce quintozone data set. For this last data set, the additional variability explained using the more complex set of thresholds was substantial, with a dramatic change in mean residue levels between years (a response that can apparently be explained by the removal of this pesticide from use towards the start of the period covered by the samples).

Secondary data

Some secondary data on pesticide usage was obtained from the Pesticide Usage Survey Reports produced by the Central Science Laboratory. However, these reports only provide a “snapshot” of pesticide usage in the years of the surveys, and so cannot be used to assess

whether detected trends in levels of pesticide residues are associated with changes in patterns of usage.

Given the relatively short time period covered by the pesticide residue data sets, and given the year-to-year and location-to-location variability in pest, disease and weed pressure, it is unlikely that any strong association could be identified between pesticide residue levels and pesticide usage and pest pressure information, even with more detailed annual records of pesticide usage and pest pressure.

Ideally, detailed information on pesticide usage and pest pressure would need to be collected either alongside the pesticide residue data, or in such a way as to allow a valid association of the residue level information with information about pesticide usage and pest pressure. Data would also be needed over a reasonably long time frame to allow a strong relationship between pesticide usage, pest pressure and residue levels to be identified, or associated with distinctly different pesticide usage practices over a wide geographical region.

Power Analysis

A power analysis was conducted to assess the ability of the new analysis method to detect shifts in the distribution of residue values. The level of resources built into the project constrained this analysis to just assess shifts in the mean residue level, but it would also be interesting to assess the power of the method to detect for changes in the variability of residue values, both in the presence and absence of shifts in the mean residue level.

Rather than using data based directly on the observed pesticide residues analysed in the first part of the project, the power analysis used data simulated from Gamma distributions. Whilst no formal assessment was made of the suitability of a Gamma distribution to describe the distribution of pesticide residue data, such distributions do have similar characteristics to the observed data (generally skewed distributions, with a large proportion of observations closer to zero than the distribution mean, and with a few larger observations).

The power analysis considered the effect of sample size, limit of determination, number of thresholds, and mean and variance of residues on the power of the new analysis method to detect differences in the distribution of pesticide levels, analysing 1000 sets of simulated data for each combination of parameters.

Results of the power analysis indicated that for relatively small sample sizes (20 – 100) the new analysis method had a tendency to detect a higher proportion of significant differences between pairs (or triples) of samples than would be expected when samples were drawn from the same underlying Gamma distribution, particularly when the limit of determination was small. When comparing more than two samples, the method tended to detect a lower proportion of significant effects than would be expected for larger sample sizes (150 – 200). So with relatively small sample sizes there is a danger that the new analysis method will detect differences between samples even if none exist, particularly when most of the observations are above the limit of determination. However, when comparing multiple samples with larger sample sizes, the new analysis method is more conservative, being unlikely to falsely detect differences between samples.

As anticipated, the power of the analysis method to detect shifts in the mean residue level depended on the mean and variance of the base-line Gamma distribution, on the sample size,

and on the limit of determination/number of thresholds used. Where the mean residue level for the base-line sample was reasonably above the limit of determination, the choice of the limit of determination was relatively unimportant. However, for comparisons involving base-line samples with a lower mean residue level, reducing the limit of determination resulted in a significant improvement of the power of the analysis method to detect differences between samples.

The results for the effect of sample size broadly followed prior expectations, with, as for many statistical tests, the power of the analysis method increasing as the sample size was increased. Again, the impact of increasing the sample size on the power of the method was less dramatic where the mean residue level for the base-line sample was higher – even with quite a small sample size, a relatively small change in mean residue level could be detected, and increasing the sample size did not result in much of a reduction in the size of difference that could be detected.

The power analysis was completed with a brief exploration of the power of the method to detect differences between multiple (more than two) samples, showing that the method was equally able to detect differences where two samples both had different means from a base-line sample, as long as both means were reasonably different from the base-line mean. However, further exploration is needed of these multiple sample scenarios, particularly with regards to detecting particular patterns (e.g. linear trends over time, or “step” changes at a particular time) of changes in mean residue levels.

Extensions

The data sets analysed in this pilot study provided some opportunities to explore the extent to which this novel statistical analysis could be used. In the original aphid distribution paper, the method was used to consider the main effects of different insecticides, different resistant populations and different population development times, as well as the interaction between these three factors.

Similarly, the apple and strawberry data sets allowed the assessment of the main effects of both year and source (UK, Imported) and the interaction between these two factors.

However, the method can be further extended to answer more specific hypotheses about differences where these are appropriate for the data set being analysed. So, for example, in the aphid distribution paper, the difference between the insecticides was partitioned into a comparison between the untreated and any insecticide, and then between insecticides with different modes of action.

So for pesticide residue data, it would be possible to group data for particular sets of years, or different sources of material and ask more specific questions about how the distribution of residue values changed between these groups.

Another extension would be to consider more formally the choice of threshold sets to categorise the values (both the number of categories, and the choice of the thresholds between categories). This will be of most relevance where the majority of residue values are greater than the LOD, and a formal comparison would also require the different threshold sets to be nested in some way (e.g. for one set to have the same thresholds as another, except for the addition of an extra intermediate threshold, thus producing an additional category) – the

comparison of the single LOD threshold analysis with the multiple threshold analysis provides an extreme example of such nesting.

However, the exploration of these extensions was beyond the resources of this project.

In addition there is the need to further extend the power analysis for this novel analysis approach as indicated above, but beyond the resources of this project.

Conclusions

- The new method does provide a potential approach for assessing for changes in pesticide residue levels, but the applicability of the method depends on a number of key characteristics of the available data:
 - Where a high proportion of the samples do not have detectable residues, the new method is unlikely to add any treatment discrimination beyond that which can be provided by a simpler analysis of the proportions of samples above and below the LOD;
 - If determination levels can be improved (lowered), then the new method is likely to be of greater value in detecting changes
 - Similarly, where the maximum observed residue level is within an order of magnitude of the LOD, there is unlikely to be sufficient variation in the distribution of residues for the new method to provide any additional discrimination over the simpler “above/below LOD” model;
- The power analysis has identified that, assuming that a Gamma distribution can be used to describe the distribution of pesticide residues, the new method can detect fairly small changes in mean residue level with sample sizes of less than 100, as long as the limit of determination is smaller than the minimum mean residue level of the samples being compared
 - As would be anticipated, larger sample sizes provide a more powerful discrimination between samples, but given more detail of the expected mean residue levels and changes in residue levels that it would be important to detect, an appropriate sample size could be determined for future sampling scenarios.
- Further extensions of the analysis method are possible to address more specific hypotheses about differences between samples (assessment of the main effects of both year and source were explored for the apple strawberry dataset in this study), and to address the choice of the threshold set used to categorise the data (extension of the power analysis would allow some exploration of this latter extension).

Detailed Description of Project

Background

In a recent research project concerned with the impact of insecticide resistance on pest management in lettuce, a novel statistical approach was developed to provide a non-parametric comparison of the distributions of the numbers of aphids per plant between different treatments (Kift, Mead *et al.*, 2004). The conventional approach to analysing data from such experiments is to analyse the mean number of aphids per plant (or total number of aphids across a group of plants within each plot) using a log-linear model (a generalised linear model assuming a Poisson error distribution and logarithm link function). However, this conventional approach was unable to detect quite large treatment differences as being significant, primarily because of the substantial over-dispersion found in the aphid count data. The conventional analysis approach is also constrained by an assumption that the shape of the underlying distribution does not change between treatments, although the mean value is obviously allowed to (this is what is being tested in the analysis).

By contrast, the approach proposed in Kift, Mead *et al* (2004) makes no assumptions about the form of distribution followed by the data, but aims to detect differences in the shapes of the distribution between treatments. This may simply be a shift in the mean value, with other characteristics of the distribution remaining constant, or could be more complex. The analysis proceeds by classifying the observed data (the aphid counts in the original development of the methodology) into a number of non-overlapping, contiguous groups. For the aphid count data these were obtained using the quartiles (lower quartile, median, upper quartile), obtained from all observed data combined across all the treatments to be compared, to produce four groups. This classification of the observed data was then combined with the treatment combinations to produce a contingency table counting the number of observations for each treatment combination in each of the four groups (defined based on the quartiles), and this contingency table was then analysed using a log-linear analysis (a generalised linear model assuming a Poisson error distribution and logarithm link function). The terms of interest in this analysis are then the interactions between the count classifying factor and the various treatment factor terms – significant interactions indicate that the distributions of counts between the four groups are different between levels of the treatment factors being considered. Note that with this form of data it is realistic to assume that there will not be over-dispersion, so that the significance of treatment terms is assessed by comparing the deviance with the critical values of the appropriate chi-square distribution.

Interpretation of significant treatment terms in these analyses is not always straightforward, and so a graphical approach was developed alongside the analysis methodology. This compares the distributions for each treatment combination through cumulative relative frequency curves, obtained by ranking the observed values (from minimum to maximum) and then plotting each value against its rank value expressed as a proportion of the total number of values.

Preliminary discussions between Neil Kift (NFU) and Food Standards Agency staff suggested that this novel statistical approach might be used to analyse for changes in pesticide residue levels. The aims of the pilot study are:

- To assess whether this novel statistical approach can detect differences in pesticide residue distributions using test data sets for apples, strawberries and protected lettuce;

- To compare the results of this new analysis approach with results from existing analysis methods;
- To determine whether detected trends in pesticide residue distributions can be related to information about pesticide usage and pest pressure;
- To assess the power of the method to detect changes in the distribution of pesticide residues, and the influence on the power of the choice of classifying thresholds, sample size, and the proportion of residues below the limit of determination (LOD).

Data

Three sets of pesticide residue data were provided by the Food Standards Agency to be used as test datasets in this pilot study:

- Apples: data from seven years (1998 – 2004) with residue levels measured for up to 6 pesticides (bupirimate, captan, carbendazim, chlorpyrifos, diphenylamine, metalaxyl), with the source of the crop being identified as either UK, Imported (divided into EC and non-EC in 2004) or Unknown. The unknown samples were omitted from the analysed data sets. Table 1 shows the limits of determination (LOD) for each pesticide and the maximum numbers of samples analysed in each year.

Table 1: Apple pesticide residue data sets – limits of determination (LOD) for each pesticide in each year (no samples were analysed for a given pesticide if no LOD is shown), and the maximum number of samples analysed in each year, classified by the source of the sample.

Year	Limit of determination						Samples		
	bupirimate	captan	carbendazim	chlorpyrifos	diphenylamine	metalaxyl	UK	Import	Unknown
1998	0.05	0.05	0.1	0.05	0.05	0.1	17	70	9
1999	-	-	-	0.01	-	-	54	83	7
2000	-	-	0.1	0.01	-	-	36	105	3
2001	0.05	0.05	0.1	0.01	0.05	0.05	21 (19*)	42 (35*)	0
2002	0.05	0.05	0.05	0.01	0.05	0.05	28	39	0
2003	0.05	0.05	0.05	0.02	0.05	0.05	82	219	0
2004	0.05	0.02	0.05	0.02	0.05	0.05	68	76	0

*smaller number of samples for carbendazim

- Strawberries: data from four years (1999, 2000, 2001, 2004) with residue levels measured for up to 4 pesticides (bupirimate, iprodione, pyrimethanil, fenhexamid) with the source of the crop being identified as either UK or imported (divide into EC and non-EC in 2004). Table 2 shows the limits of determination (LOD) for each pesticide and the number of samples analysed in each year.

Table 2: Strawberry pesticide residue data sets – limits of determination (LOD) for each pesticide in each year (no samples were analysed for a given pesticide if no LOD is shown), and the maximum number of samples analysed in each year, classified by the source of the sample.

Year	Limit of determination				Samples	
	bupirimate	iprodione	pyrimethanil	fenhexamid	UK	Import
1999	0.02	0.05	0.05	-	27	18
2000	0.02	0.02	0.02	-	0	11
2001	0.02	0.05	0.02	0.05	62 (89*)	16 (67*)
2004	0.05	0.02	0.02	0.05	42	56

*higher numbers of samples analysed for fenhexamid

- Protected lettuce: data from 10 sampling occasions (over 5 years from 2002 – 2006) with residue levels measured for up to 9 pesticides (iprodione, propyzamide, azoxystrobin, dithiocarbamates, cypermethrin, inorganic bromide, quintozone, pyrimethanil, fenhexamid). Table 3 shows the limits of determination (LOD) for each pesticide and the number of samples analysed in each year.

Table 3: Protected lettuce pesticide residue data sets – limits of determination (LOD) for each pesticide in each year (no samples were analysed for a given pesticide is no LOD is shown), and the maximum number of samples analysed in each year.

Occasion	Limit of determination									Samples
	iprodione	propyzamide	azoxystrobin	dithiocarbamates	cypermethrin	inorganic bromide	quintozone	pyrimethanil	fenhexamid	
2002 Feb/Mar	0.01	0.01	0.01	0.05	0.05	2.0	0.01	-	-	20
2002 Nov	0.01	0.02	0.02	0.02	0.02	10.0	0.01	-	-	25
2003 Feb/Mar	0.01	0.02	0.02	-	0.02	-	0.01	-	-	28
2003 Oct	0.01	0.02	0.02	-	0.02	10.0	-	-	-	21
2004 Feb/Mar	0.01	0.02	0.02	-	0.02	10.0	0.01	-	-	26
2004 May	0.01	0.02	0.02	0.02	0.02	10.0	0.01	-	-	32
2004 Oct/Nov	0.01	0.02	0.02	0.02	0.02	10.0	-	-	-	33
2005 Mar	0.01	0.02	0.02	0.02	0.02	10.0	0.01	-	-	29
2005 Nov/Dec	0.01	0.02	0.02	0.02	0.02	10.0	0.01	0.01	0.03	28
2006 Feb/Apr	0.01	0.02	0.02	0.05	0.02	10.0	0.01	0.01	0.03	34 (31*)

*smaller number of samples for inorganic bromide

Methodology

Initial data summary: the proportion of non-zero residues (i.e. those measured above the limit of determination), the mean non-zero residue level, and the maximum recorded residue level were calculated for each pesticide for each year (apples, strawberries) or sampling occasion (protected lettuce). Where the source of the sample was identified (apples, strawberries), an additional summary of the numbers of samples below and above the limit of determination was calculated for each source separately.

Comparison of “treatments”: analyses were performed for datasets constructed for each different pesticide within each crop.

Two simple analysis approaches were performed for each dataset to provide a base-line result against which the result from the new method could be compared:

- Analysis of variance of all residue data, classified by year and source for the apple and strawberry datasets, and by either sampling occasion or year and season (spring or autumn for the protected lettuce datasets). Whilst the ANOVA assumptions of normality and homogeneity of variance would almost certainly fail for every dataset, and the residual variance would be seriously underestimated because of the large number of zeros in most data sets, this analysis still provides a base-line of an approach that might be used. Note that the unbalanced nature of each dataset means that a regression-based approach needed to be taken, with the analysis of variance summary depending on the order in which terms were added. A non-parametric ANOVA was also considered, but the large number of repeated values (zeros) would cause the assumptions for such an analysis to also fail.
- Analysis of the proportions of residues above and below the limit of determination (LOD), by analysing a contingency table classified by “treatment” (year and source, or

sampling occasion, or year and season as appropriate) and “above/below LOD” using a log-linear model (generalised linear model assuming a Poisson error distribution and logarithm link function). After first fitting the main effect of “above/below LOD” (we are not interested in whether these proportions are equal overall) and the main effects and interactions between the “treatment” (we know that the numbers of samples will vary between year/source/sampling occasion), we then fit terms for the interactions between “treatments” and “above/below LOD”, fitting these terms in all possible orders to allow for the non-orthogonality between treatment terms. This is essentially the simplest version of the new approach being investigated, with just one “threshold” value, the LOD, used to categorise the observed residues. Note that where the LOD varied between datasets, the observed values were re-evaluated relative to the highest LOD so that the proportional response relative to a consistent LOD was obtained. Where a relatively large number of “treatments” were observed relative to a lower LOD, a secondary analysis of the appropriately restricted dataset was also performed.

The new analysis approach was similarly applied to each of the pesticide datasets for each crop. A potentially important decision for each analysis, to be explored further in the power analysis element of the study, is the choice of thresholds above the LOD to be used to classify the observed residue values. For most datasets a simple, generic, approach was used, choosing thresholds based on multiples of the LOD, following a logarithmic-type scale – so, for example, with an LOD of 0.01, higher thresholds of 0.05 (LOD * 5), 0.10 (LOD * 10), 0.50 (LOD * 50), 1.00 (LOD*100), etc., might be used. Depending on the range of observed residue levels, some of the intermediate thresholds (e.g. 5 * LOD, 50* LOD) might be omitted to ensure a reasonable number of observations in each category. For a couple of datasets with a relatively large proportion of residues above the LOD, an alternative approach, based on that used originally for the aphid number distributions was also tried – quartiles were calculated based on all the non-zero residue values, and the observed non-zero values classified into four groups defined by the quartiles in addition to the “below the LOD” group.

The analysis was performed using a log-linear model (generalised linear model assuming a Poisson error distribution and logarithm link function). As for the simpler “above/below LOD” version of the approach, as described above, we first fit the main effect of residue classification and the main effects and interactions between the “treatments”, and then assess the significance of the interactions between the residue classification and “treatments”, fitting terms in all possible orders to allow for the non-orthogonality between treatment terms.

For each dataset, cumulative relative frequency plots were produced to aid interpretation. For the apple and strawberry datasets, curves were constructed for each year (combined across sources) and for each source separately within each year. For the protected lettuce dataset, curves were produced for each sampling occasion.

A second visual interpretation tool was developed within this project, using stacked barcharts to indicate the proportions of observations in each of the classification groups. In addition to providing a simple visual summary of the differences between the distributions for different sampling occasions, this tool also provides a graphical comparison of the single and multiple threshold sets.

A formal assessment of the additional discrimination provided by the extended new analysis method (with multiple classification groups above the LOD) compared to the simpler approach (above/below LOD) was obtained by considering the additional variability

explained by the more complex threshold set. This can be obtained by calculating the difference in the deviances for each interaction term between the analysis for the more complex threshold set and that for the single threshold (LOD) set. A test of whether the more complex threshold set adds any further discrimination can then be obtained by comparing this extra deviance with the critical values of the appropriate chi-square distribution.

All data summaries and analyses were performed using GenStat for Windows. Cumulative relative frequency graphs and stacked bar-charts were constructed using Excel.

Results

Initial Data Summary

- Apples:
 - bupirimate – LOD consistent across years, but maximum residue only twice LOD, and a maximum of 3.2% of residues above the LOD for any year;
 - captan – lower LOD in final year (re-evaluation of data to higher LOD), with maximum residue from 2 x max(LOD) to 16 x max(LOD), and between 6% and 29% of residues above the LOD in any year (highest in final year when LOD lower!);
 - carbendazim – lower LOD for final three years (of six), with maximum residue from 3 x max(LOD) to 12 x max(LOD), and between 10% and 24% of residues above the LOD;
 - chlorpyrifos – three different LODs, with maximum residue reasonably high compared with lowest LOD but fairly close to highest LOD, and with 20% to 40% of residues above lower LODs;
 - diphenylamine – LOD consistent across years, with maximum residue substantially greater than LOD, and 23% to 46% of residues above the LOD;
 - metalaxyl – higher LOD in first year, but maximum residues not much higher than LOD, and a maximum of 6% of residues above the LOD in any year (highest when LOD lowest).
 - The new analysis method may be of value for discriminating between years and sources for captan, chlorpyrifos (restricted dataset), diphenylamine, and possibly carbendazim (restricted dataset).
- Strawberries:
 - bupirimate – higher LOD in final year, but with relatively high maximum residues relative to max(LOD) – around 30% of residues above lower LOD, but only around 6% for higher LOD;
 - iprodione – two LOD levels for two years each (need to re-evaluate residues relative to higher LOD), but with maximum residues substantially higher than max(LOD), and 27% to 49% of residues above LOD (highest for higher LOD!);
 - pyrimethanil – higher LOD in first year, but with relatively high maximum residues relative to max(LOD), and 15% to 50% of residues above the LOD;

- fenhexamid – consistent LOD for two years of samples, with maximum residues substantially higher than LOD, and 20% to 37% of residues above the LOD.
- The new analysis method may be of value for discriminating between years and sources for all four pesticides.
- Protected lettuce:
 - iprodione – consistent LOD, with maximum residue substantially higher than LOD on all occasions, and between 6% and 85% of residues greater than the LOD;
 - propyzamide – lower LOD on first occasion, with maximum residue very variable between occasions, and between 0% (2 occasions) and 46% of residues above the LOD;
 - azoxystrobin – lower LOD on first occasion, with maximum residue very variable between occasions, and between 0% (2 occasions) and 28% of residues above the LOD;
 - dithiocarbamates – higher LOD on first and last occasions (of seven), with maximum residue usually substantially higher than the LOD, and between 18% and 30% of residues above the LOD;
 - cypermethrin – higher LOD on first occasion, with maximum residue usually substantially higher than the LOD, and between 9% and 57% of residues above the LOD;
 - inorganic bromide – lower LOD on first occasion, with maximum residue very variable between occasions, but usually substantially higher than LOD, and between 23% and 40% of residues above the LOD;
 - quintozene – consistent LOD, with between 0% and 75% of residues above the LOD, but with maximum residue usually only a little higher than LOD (2- or 3-times the LOD);
 - pyrimethanil and fenhexamid – consistent LODs (only two occasions each), with maximum residues substantially higher than LOD, but only 12% to 22% of residues above the LOD.
 - New analysis method may be of value for discriminating between occasions for all pesticides, though probably less promising for quintozene than for other pesticides.

Results from New Approach

- Results of the analyses of the test datasets are summarised below, by chemical within crop, with a few example results shown in more detail.
- Apples:
 - Captan – analysis indicated significant effects of both year and source on the distribution of residues, but no effect of the interaction between year and source;
 - Carbendazim – analysis indicated significant effects of both year and source on the distribution of residues, and also a significant effect on the distribution of residues of the interaction between year and source (for two different sets of classification thresholds);

- Table 4 shows the distributions of carbendazim residues across the three selected ranges for each of the year by course data sets

Table 4: Carbendazim residues in apples – total sample size for UK and imported samples in each of the six years, with numbers and percentages (in red) of samples in each of three ranges

Year	Source	Sample size	Range		
			Below LOD (0.1 mg/kg)	0.1 – 0.3 mg/kg	> 0.3 mg/kg
1998	UK	17	12 70.6 %	3 17.6 %	2 11.8 %
	Import	70	56 80.0 %	13 18.6 %	1 1.4 %
2000	UK	36	24 66.7 %	5 13.9 %	7 19.4 %
	Import	105	93 88.6 %	5 4.8 %	7 6.7 %
2001	UK	19	18 94.7 %	0 0.0 %	1 5.3 %
	Import	35	31 88.6 %	2 5.7 %	2 5.7 %
2002	UK	28	20 71.4 %	6 21.4 %	2 7.1 %
	Import	39	37 94.9 %	1 2.6 %	1 2.6 %
2003	UK	82	72 87.8 %	6 7.3 %	4 4.9 %
	Import	219	202 92.2 %	15 6.8 %	2 0.9 %
2004	UK	68	47 69.1 %	9 13.2 %	12 17.6 %
	Import	76	74 97.4 %	2 2.6 %	0 0.0 %

- Differences between years predominantly seen in the higher percentage of samples below the LOD in years 2001 and 2003, most notably in the UK samples; differences between sources predominantly seen with higher percentages below the LOD for Imported samples;
- As these differences are predominantly associated with the percentage of samples below the LOD, the cumulative relative frequency graphs (Figures 1, 2 and 3) do not provide a particularly clear interpretation, although the curves for 2001 (green) and 2003 (dark blue) can be seen at the top of the set of curves in both the overall (Figure 1) and UK (Figure 2) graphs;

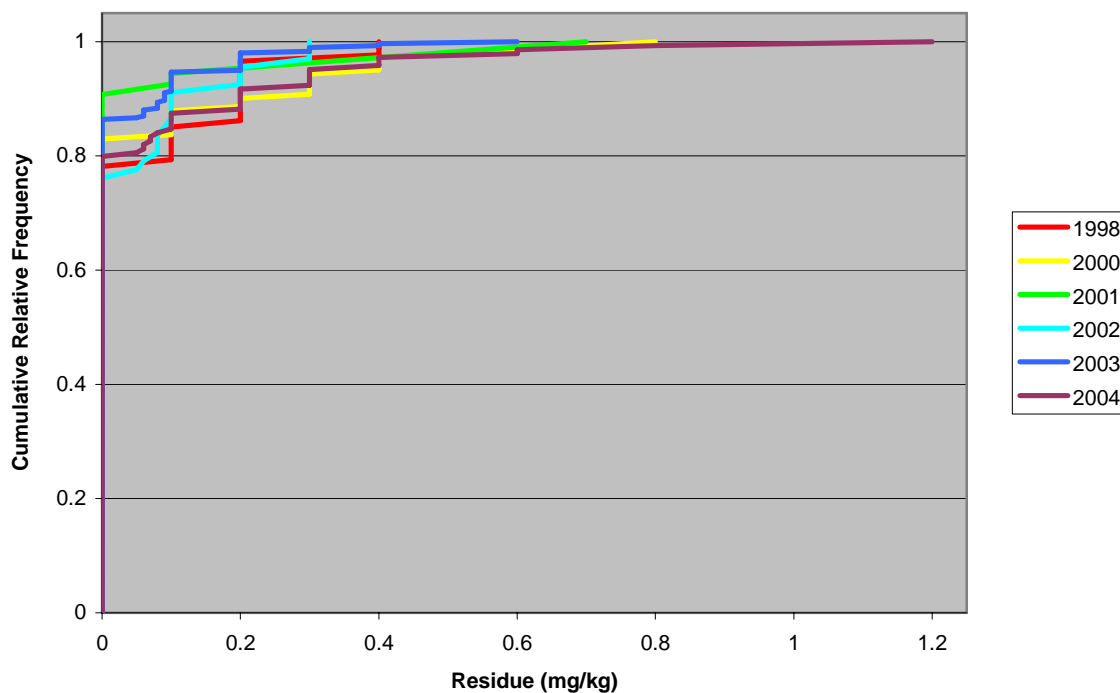


Figure 1: Distribution of carbendazim residues in apples (UK and imported combined) for each of six years – cumulative relative frequency plots showing the proportion of samples below each residue level.

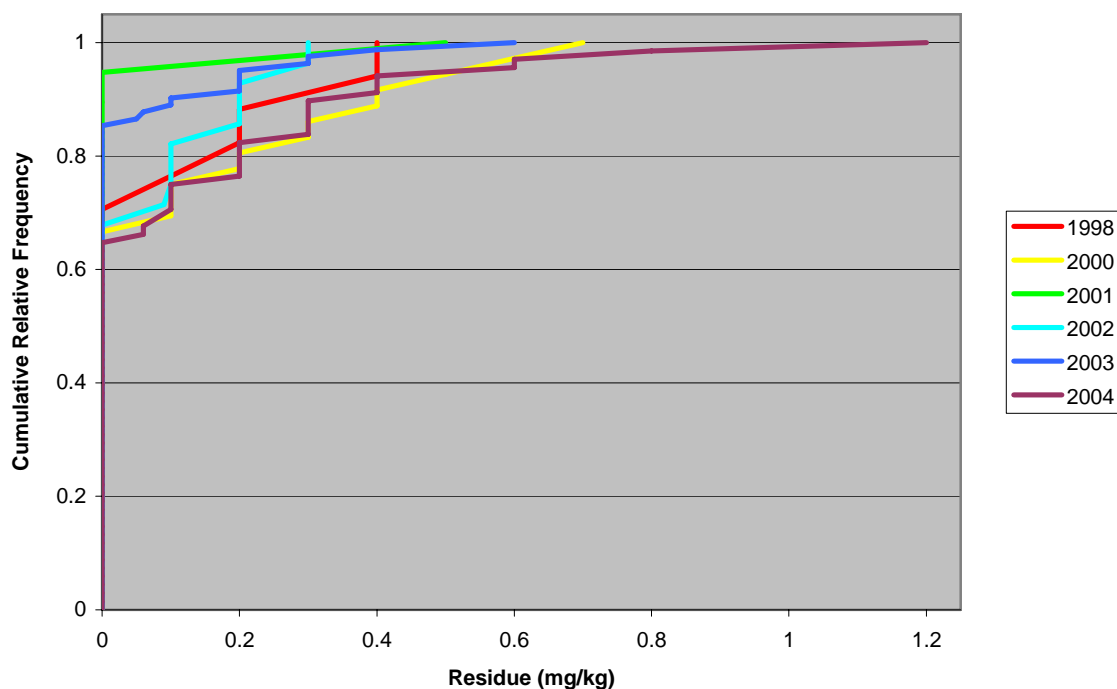


Figure 2: Distribution of carbendazim residues in UK apples for each of six years – cumulative relative frequency plots showing the proportion of samples below each residue level.

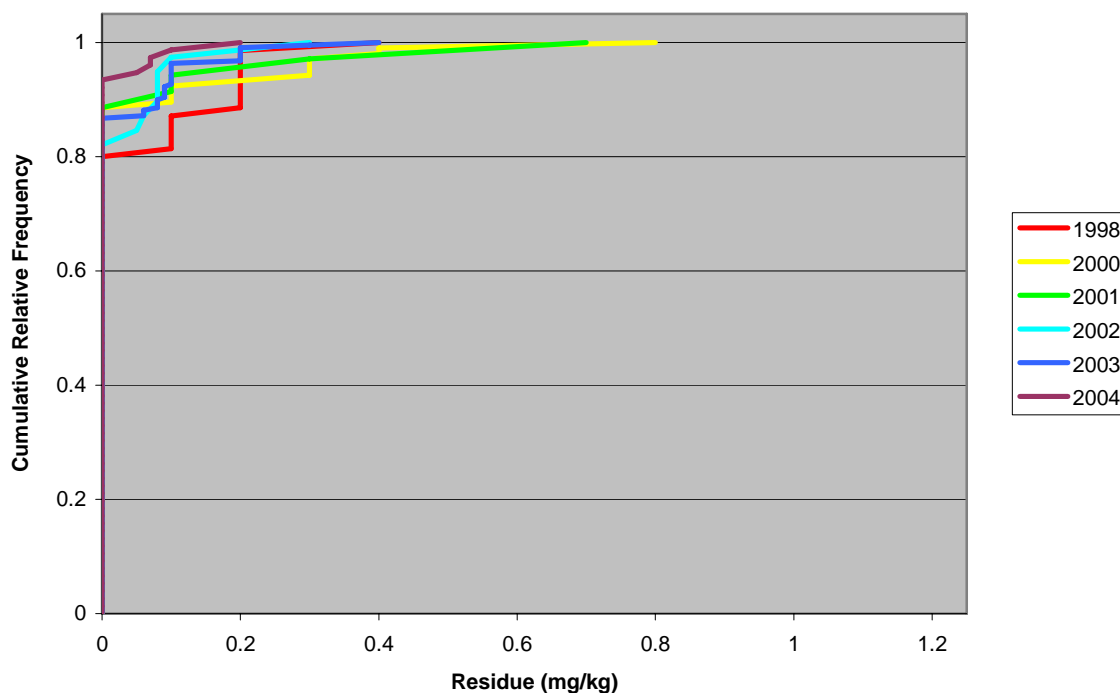


Figure 3: Distribution of carbendazim residues in imported apples for each of six years – cumulative relative frequency plots showing the proportion of samples below each residue level.

- The higher percentage of Imported samples with residues below the LOD can, however, be clearly seen from the comparison of UK and Imported graphs;
- An alternative graphical presentation of the distribution of values between the different ranges is given in the stacked bar-charts (Figure 4). The length of each bar indicates the percentage of residues in that range, with the light blue bar showing the percentage of residues below the LOD. As this is the bar that changes most, this gives an indication that there is relatively little to be gained in this case from using the more complex, multiple threshold set;

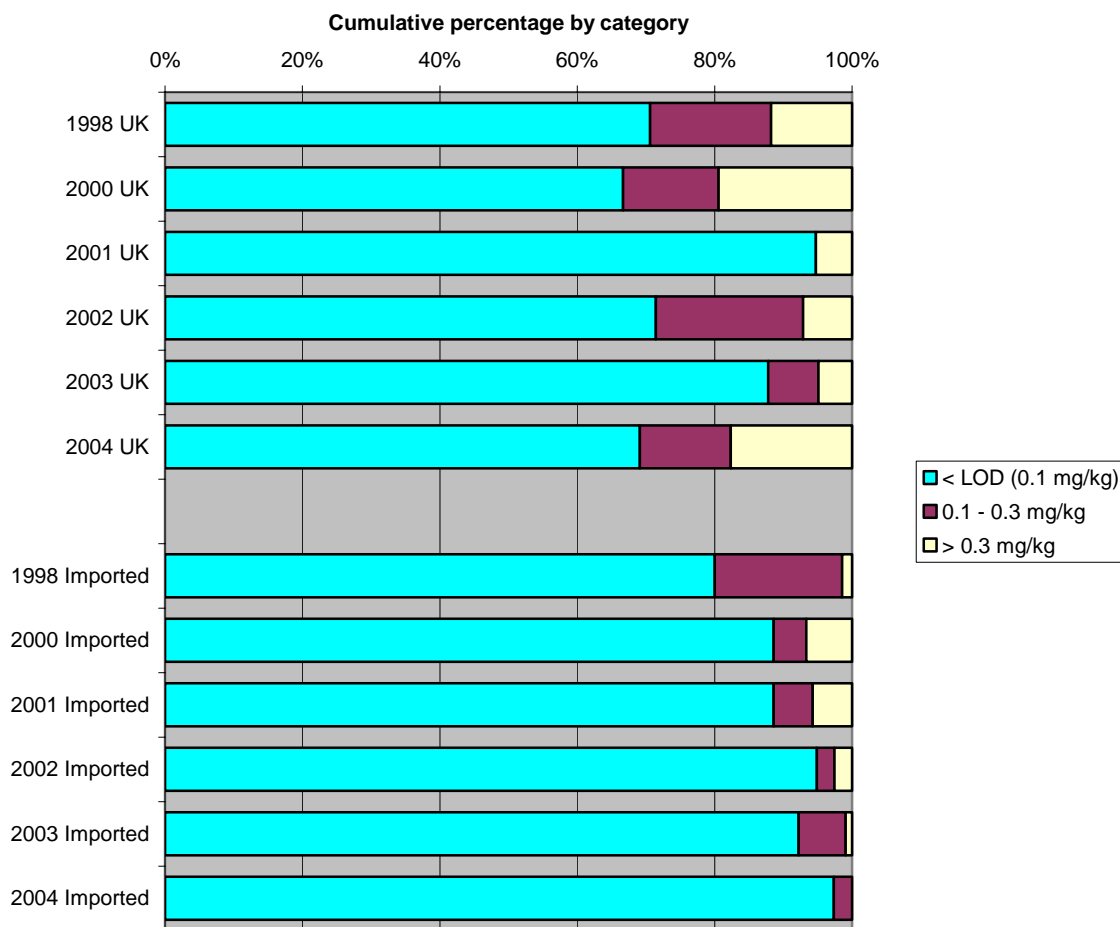


Figure 4: Distribution of carbendazim residues in apples for UK and imported samples in each of six years – stacked percentage bar charts showing the (cumulative) percentage of samples in each of three residue ranges

- Chlorpyrifos – analysis indicated a significant effect of source on the distribution of residues, but no significant effect of either year or the interaction between year and source (assessed for two different sets of classification thresholds);
 - Analysis using the quartiles of the non-zero residues to define the classification thresholds gave a stronger indication of year differences;
 - Table 5 shows the counts of residues in each of the five ranges based on the quartile classification, with the counts also expressed as percentages of the total sample for each year by course combination;
 - The striking difference here is in the higher proportions of higher residues for the UK samples, particularly in 2002 (the average residue level seems to be higher in this year and the percentage of residues below the LOD was lower than for the other years);
 - This interpretation is displayed graphically in Figure 5;

Table 5: Chlorpyrifos residues in apples – total sample size for UK and imported samples in each of the six years, with numbers and percentages (in red) of samples in each of five ranges defined by the quartiles of the complete data set.

Year	Source	Sample size	Below LOD (0.02)	LOD – P ₂₅	P ₂₅ – P ₅₀	P ₅₀ – P ₇₅	> P ₇₅
1999	UK	54	34 63.0%	5 9.3%	1 1.9%	9 16.7%	5 9.3%
	Import	83	71 85.5%	5 6.0%	3 3.6%	3 3.6%	1 1.2%
2000	UK	36	17 47.2%	3 8.3%	4 11.1%	6 16.7%	6 16.7%
	Import	105	86 81.9%	6 5.7%	6 5.7%	5 4.8%	2 1.9%
2001	UK	21	11 52.4%	4 19.0%	1 4.8%	3 14.3%	2 9.5%
	Import	42	39 92.9%	3 7.1%	0 0.0%	0 0.0%	0 0.0%
2002	UK	28	13 46.4%	2 7.1%	1 3.6%	6 21.4%	6 21.4%
	Import	39	29 74.4%	2 5.1%	3 7.7%	3 7.7%	2 5.1%
2003	UK	82	41 50.0%	7 8.5%	9 11.0%	13 15.9%	12 14.6%
	Import	219	198 90.4%	3 1.4%	5 2.3%	8 3.7%	5 2.3%
2004	UK	68	42 61.8%	2 2.9%	3 4.4%	11 16.2%	10 14.7%
	Import	76	71 93.4%	1 1.3%	0 0.0%	2 2.6%	2 2.6%

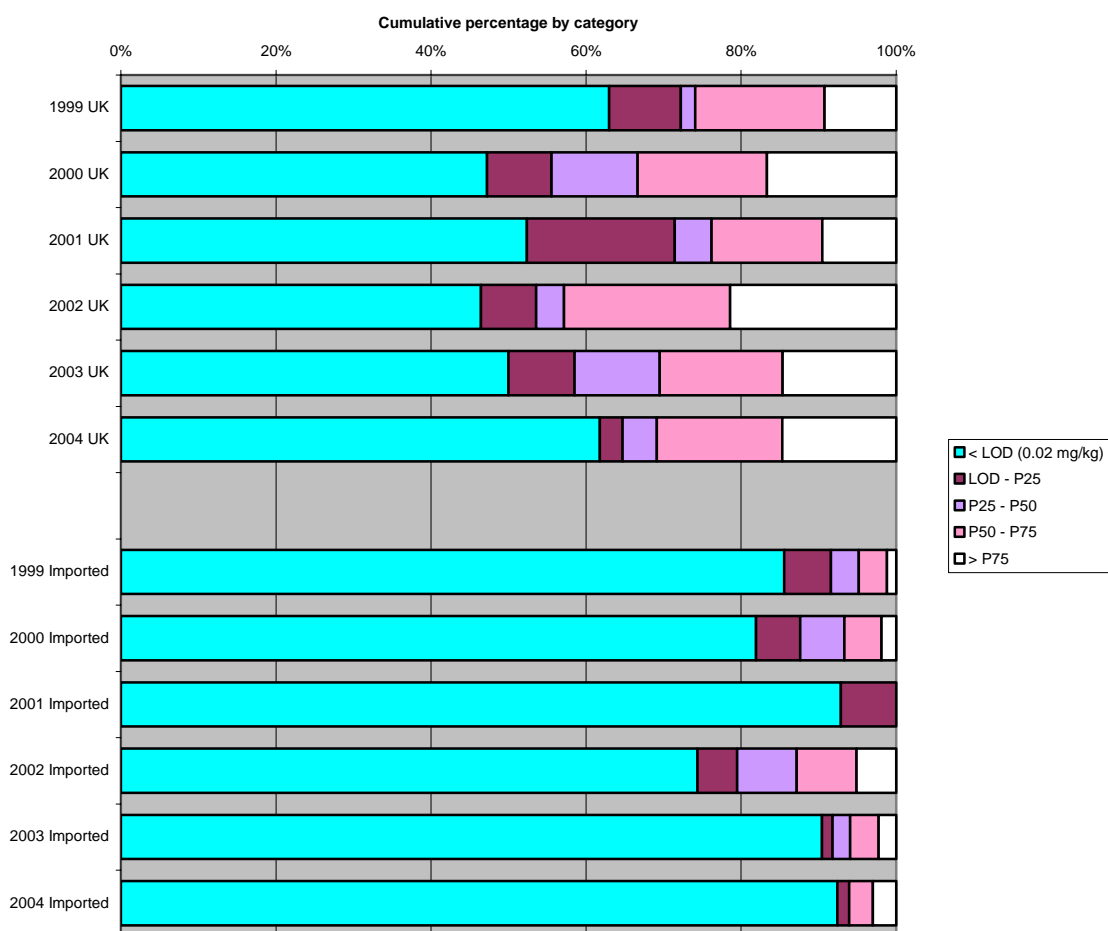


Figure 5: Distribution of chlorpyrifos residues in apples for UK and imported samples in each of six years – stacked percentage bar charts showing the (cumulative) percentage of samples in each of five residue ranges

- Diphenylamine – analysis indicates a significant effect of source and a significant effect of the interaction between year and source, but no significant effect of year;
 - Analysis using the quartiles of the non-zero residues to define the classification thresholds did not give any improved discrimination between years for this pesticide;
- Bupirimate/metalaxyl – no analysis performed as there were too few non-zero residues.
- Strawberries
 - Bupirimate – analysis of all years, with data re-evaluated against the higher LOD (0.05), indicates a significant effect of source on the distribution of residues, but no significant effect of year or of the interaction between year and source;
 - Analysis of the restricted dataset (excluding 2004 data) based on the lower LOD (0.02) indicates no significant effects
 - Iprodione – analysis indicates a significant effect of year, a possibly significant effect of source (depends on the order in which terms are added to the model), and a significant effect of the interaction between year and source;
 - Pyrimethanil – analysis of all years, with data re-evaluated against the higher LOD (0.05), indicates significant effects of both year and source, but no significant effect of the interaction between year and source;
 - Analysis of the restricted dataset (excluding 1999 data) based on the lower LOD (0.02) shows similar patterns;
 - Fenhexamid – analysis indicates a significant effect of source, possibly a significant effect of year (depends on the order in which terms are added to the model), but no significant effect of the interaction between year and source.
- Protected lettuce:
 - Iprodione - analysis shows a highly significant effect of occasion on the distribution of pesticide residues;
 - Table 6 shows the counts of residues in each category for each sampling occasion, together with these values as percentages of the total sample size for each sampling occasion;
 - The treatment effect is predominantly seen in the greater spread of residues for the assessments made in February and March, with a lower percentage of residues below the LOD and a higher percentage above 1.0 mg/kg;

Table 6: Iprodione residues in protected lettuce – total sample size for each of the ten sampling occasions, with numbers and percentages (in red) of samples in each of five ranges.

Occasion	Sample size	Below LOD (0.01)	0.01 – 0.1	0.1 – 1.0	1.0 – 10.0	> 10.0
2002 Feb/Mar	20	4 20.0%	5 25.0%	6 30.0%	5 25.0%	0 0.0%
2002 Nov	25	18 72.0%	3 12.0%	2 8.0%	2 8.0%	0 0.0%
2003 Feb/Mar	28	13 46.4%	4 14.3%	5 17.9%	5 17.9%	1 3.6%
2003 Oct	21	17 81.0%	2 9.5%	2 9.5%	0 0.0%	0 0.0%
2004 Feb/Mar	26	4 15.4%	7 26.9%	11 42.3%	4 15.4%	0 0.0%
2004 May	32	30 93.8%	0 0.0%	2 6.3%	0 0.0%	0 0.0%
2004 Oct/Nov	33	20 60.6%	7 21.2%	4 12.1%	2 6.1%	0 0.0%
2005 Mar	29	9 31.0%	10 34.5%	5 17.2%	5 17.2%	0 0.0%
2005 Nov/Dec	28	19 67.9%	4 14.3%	1 3.6%	4 14.3%	0 0.0%
2006 Feb/Apr	34	15 44.1%	8 23.5%	9 26.5%	2 5.9%	0 0.0%

- These differences can be seen in the cumulative relative frequency plot (Figure 6), with the lines for the February/March samples (red, orange, green, purple) generally being towards the bottom of the set of curves, and finishing further to the right. The two samples which are exceptions to this general pattern are those from 2005 November/December (lilac) and 2006 February/March/Apr (black)

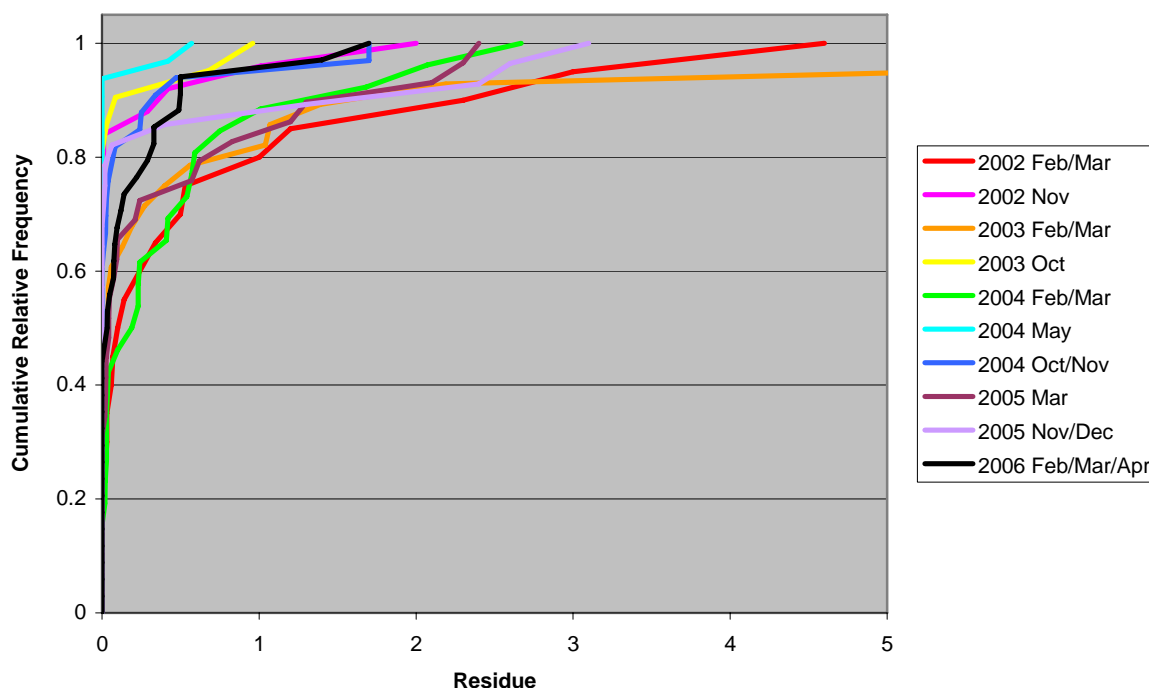


Figure 6: Distribution of iprodione residues in protected lettuce for each of ten sampling occasions – cumulative relative frequency plots showing the proportion of samples below each residue level.

- The stacked bar-chart (Figure 7) again possibly gives a clearer picture of the differences between distributions for the different sampling occasions. The light blue bars again indicate the percentage of residues below the LOD, and whilst these bars change in length most dramatically, it is likely that the extra discrimination provided by using the more complex multiple threshold set is of benefit here;

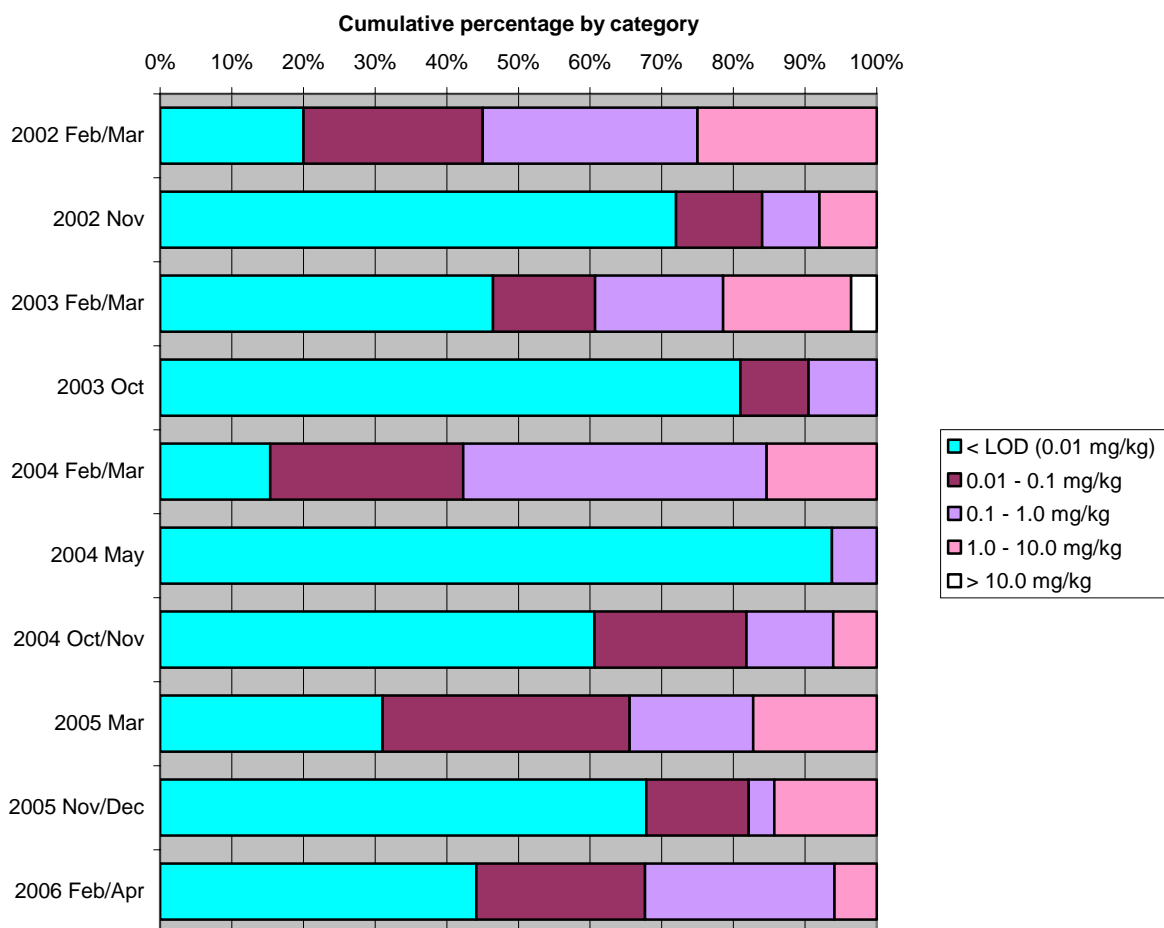


Figure 7: Distribution of iprodione residues in protected lettuce samples on each of ten occasions – stacked percentage bar charts showing the (cumulative) percentage of samples in each of five residue ranges.

- Propyzamide – analysis shows a highly significant effect of occasion on the distribution of residues;
- Azoxystrobin – analysis shows no significant effect of occasion;
- Dithiocarbamates – analysis shows no significant effect of occasion;
- Cypermethrin – analysis of full data set, with data re-evaluated against the higher LOD (0.05) shows a significant effect of occasion on the distribution of residues;
 - Analysis of restricted dataset (excluding 2002 Feb/Mar data) based on the lower LOD (0.02) also shows a significant effect of occasion on the distribution of residues;

- Inorganic bromide – analysis shows no significant effect of occasion;
- Quintozene – analysis shows a highly significant effect of occasion on the distribution of residues;
- Pyrimethanil – analysis shows no significant effect of occasion (data only available for two occasions);
- Fenhexamid – analysis shows no significant effect of occasion (data only available for two occasions).

- Further analysis attempted to disentangle the effects of year and time of year – 2004 May assessment makes it difficult to do this including all data, particularly as this occasion often has a different distribution to the February/March assessments in different years, thus not allowing the separation into “spring” and “autumn” samples.

Comparisons with Existing Approaches

- ANOVA results generally suggested stronger “treatment” differences than the new method, though this is not completely surprising due to the reduced residual mean square as a result of the often large number of zero observations.
- Similarly, analysis of the numbers of samples above and below the LOD (data sets as indicated above) also generally showed stronger “treatment” differences than the new method;
 - The results of these analyses indicated that there were often quite large changes in residue distributions that could be detected purely on the basis of the proportion of residues above and below the LOD;
 - This method might therefore be used as a simple approach to the detection of important changes in residue levels between years, or as a result of policy changes, where the observed data are not sufficient for the new method to be used (see below).
- As the “above/below LOD” model is essentially a simplification of the new method, a direct comparison of the two models for explaining differences in the data can be obtained by considering the change in deviance between the two models (see Methodology section above). These formal comparisons are summarised below:
- Apples:
 - Captan – evidence for additional discrimination provided by the new model over the “above/below LOD” model for the effect of source;
 - Carbendazim – evidence for additional discrimination for the effects of source and year for one set of classification thresholds, but not for the other;
 - Chlorpyrifos – evidence for additional discrimination for all effects for the “quartiles” based classification thresholds, but not for the other classification threshold sets;
 - Diphenylamine – evidence for additional discrimination for the effect of source for the “quartiles” based classification thresholds, but no evidence for the other set of classification thresholds.

- Strawberries:
 - Bupirimate, pyrimethanil, fenhexamid – no evidence for any additional discrimination provided by the new model over the “above/below LOD” model;
 - Iprodione – evidence for additional discrimination provided by the new model over the “above/below LOD” model for the effect of year.

- Protected lettuce:
 - Iprodione, propyzamide, azoxystrobin, dithiocarbamates, cypermethrin, pyrimethanil – no evidence for any additional discrimination provided by the new model over the “above/below LOD” model;
 - Inorganic bromide, fenhexamid – borderline evidence for additional discrimination for the effect of occasion.
 - Quintozene – evidence for additional discrimination provided by the multiple threshold set over the “above/below LOD” single threshold for the effect of occasion;
 - Table 7 shows the counts of residues for both threshold sets:
 - The “Below LOD” and “Above LOD” columns show the counts of residues in the two categories for the single threshold analysis;
 - The “Below LOD” column and the three right hand columns show the counts of residues in four categories for the multiple threshold set analysis;
 - In both cases, the red values are these counts expressed as a percentage of the total sample size;

Table 7: Quintozene residues in protected lettuce – total sample size for each of the eight sampling occasions, with numbers and percentages (in red) of samples in each of five ranges – the “Above LOD” category is sub-divided into categories based on additional thresholds to demonstrate the additional information provided by the new analysis method.

Occasion	Sample Size	Below LOD (0.01)	Above LOD	0.01 – 0.02	0.02 – 0.05	> 0.05
2002 Feb/Mar	20	6 30.0%	14 70.0%	0 0.0%	0 0.0%	14 70.0%
2002 Nov	25	14 56.0%	11 44.0%	5 20.0%	6 24.0%	0 0.0%
2003 Feb/Mar	28	7 25.0%	21 75.0%	4 14.3%	16 57.1%	1 3.6%
2004 Feb/Mar	26	14 53.8%	12 46.2%	11 42.3%	1 3.8%	0 0.0%
2004 May	32	31 96.9%	1 3.1%	1 3.1%	0 0.0%	0 0.0%
2005 Mar	29	27 93.1%	2 6.9%	2 6.9%	0 0.0%	0 0.0%
2005 Nov/Dec	28	28 100.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%
2006 Feb/Apr	34	33 97.1%	1 2.9%	1 2.9%	0 0.0%	0 0.0%

- the extra information provided by the multiple classification groups is fairly obvious, with the number of high residues (and the mean detectable residue) gradually reducing from 2000 to 2004, with most residues being undetectable in 2005 and 2006;
- This pattern can be seen graphically using the stacked bar-chart approach (Figure 8).

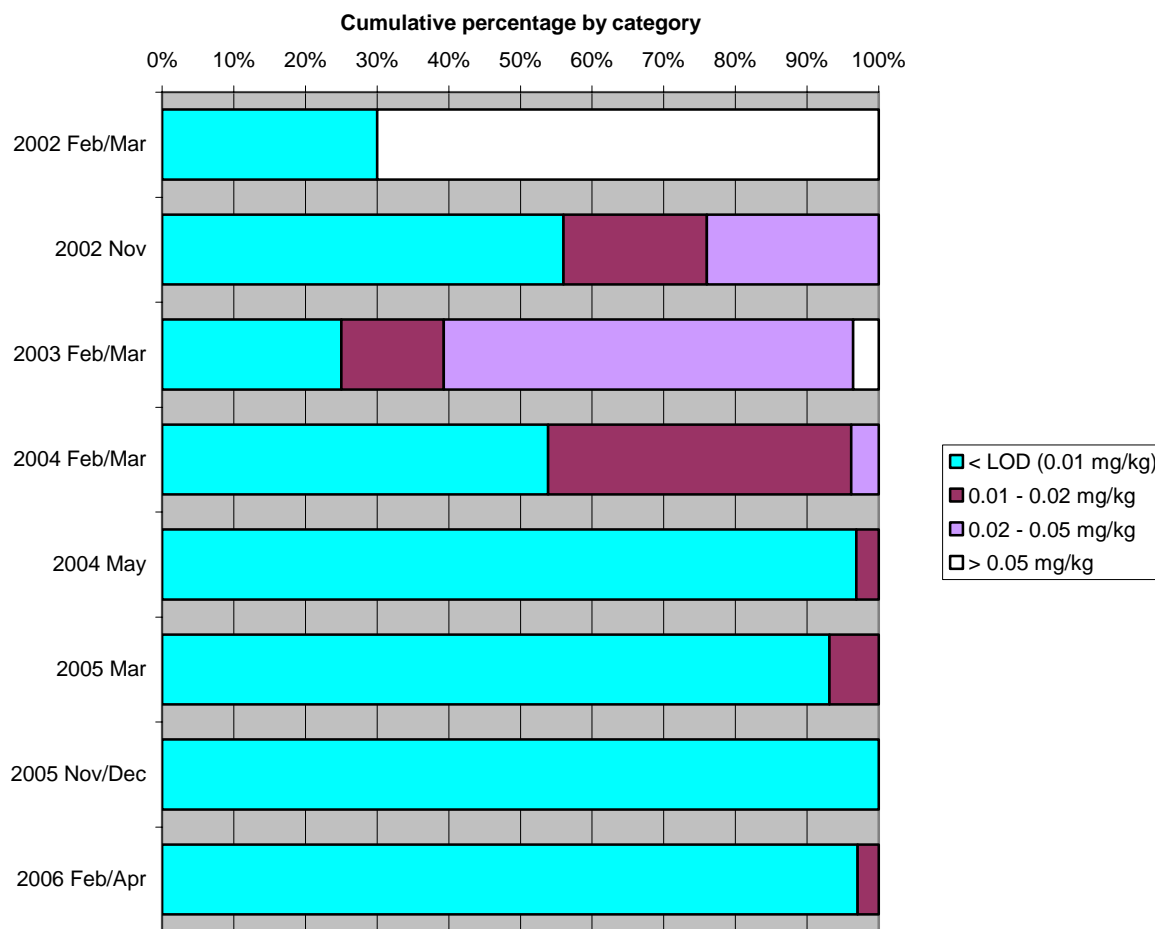


Figure 8: Distribution of quintozene residues in protected lettuce samples on each of eight occasions – stacked percentage bar charts showing the (cumulative) percentage of samples in each of four residue ranges.

Limitations of the approach

For a number of the data sets analysed in this pilot study, a number of the cells of the contingency table contained counts either close to or equal to zero. Whilst the log-linear analysis method will generally cope with such observations (occasionally having problems in convergence of the iterative method used to fit the log-linear model), these cells with both small observed counts and, potentially, small expected counts will generally have the effect of reducing the power of the analysis (they contribute degrees of freedom to the interaction term without contributing much to the deviance). This may result in the analysis not detecting real changes in the distribution of residues. This limitation is explored further in the power analysis (see below).

Secondary Data

It appears to be relatively difficult to obtain comprehensive data on pesticide usage and pest/disease/weed pressure for all years covered by the data sets considered in this pilot study. Tables 8, 9 and 10 give information for each of the pesticides included in the apple and strawberry data sets from recent Pesticide Usage Survey Reports produced by the Central Science Laboratory. Obviously these only give a “snapshot” of the usage in the particular years of survey.

To obtain more detailed information would probably mean surveying individual growers to gather information both on the pesticides used and the levels of pest/disease/weed pressure that have triggered the use of these pesticides. Ideally, this information should be gathered alongside the pesticide residue samples, so that a direct correlation could be made between pesticide usage, pest pressure and the presence of pesticide residues. Probably, though, the two data collection exercises need to be kept separate but collected in such a way as to validate the direct comparison between sources of information.

Table 8: Apples – pesticide usage data for pesticides applied in orchards in the UK in 2000 and 2004 (taken from the CSL Pesticide Usage Survey Reports)

Active substance	Area sprayed		Area sprayed as % of area grown		Average no. applications where applied		Proportion of label rate		Kg active ingredient	
	2000	2004	2000	2004	2000	2004	2000	2004	2000	2004
bupirimate	15682	17582	126	181	*	*	*	*	2806	3529
captan	53180	46514	428	479	*	5.6	*	0.35	49453	45340
carbendazim	6903	27374	56	282	*	5.1	*	0.40	2013	6613
chlorpyrifos	26075	17858	210	184	*	2.4	*	0.72	15846	12171

Area of dessert and culinary apples grown: 2000 = 12432 ha, 2004 = 17582 ha

Table 9: Apples – pesticide usage data for pesticides applied post harvest in the UK in 2000 and 2004 (taken from the CSL Pesticide Usage Survey Reports)

Active substance	Tonnes treated		Area tonnes treated as % of tonnes stored		Kg active ingredient used	
	2000	2004	2000	2004	2000	2004
captan	1493	1564	1	1	17	34
carbendazim	28523	23581	26	21	485	87
carbendazim/metalaxyl	64258†	-	59	-	586†	-
diphenylamine	41060	23083	37	20	977	512
metalaxyl	-	18925	-	16	-	75

Tonnes of apples stored: 2000 = 109506 t, 2004 = 114962 t

† Always used in admixture with carbendazim in 2000 because of product formulation

Table 10: Strawberries - pesticide usage data for pesticides applied in the UK in 1998 and 2001 (taken from the CSL Pesticide Usage Survey Reports)

	Area sprayed		Area sprayed as % of area grown		Total Kg active ingredient used	
	1998	2001	1998	2001	1998	2001
bupirimate	3340	2857	86	76	1097	928
iprodione	4435	2857	114	76	2750	1923
pyrimethanil	2647	2324	68	62	1597	1459
fenhexamid	-	3207	-	85	-	2252

Area of strawberries grown: 1998 = 3887 ha, 2001 = 3765 ha

Power Analysis

Approach

The aim of the power analysis was to assess the ability of the analysis method to detect shifts in the distribution of residue values. Time constraints within the project mean that this assessment was only concerned with shifts in the mean residue level – it would also be interesting to assess the power of the method to detect changes in the variability of residues, but this was beyond the resources of this project.

Although no formal assessment was made of how well particular statistical distributions could describe the observed residue data, the Gamma distribution was considered to have appropriate characteristics (non-negative values, often highly skewed when the mean parameter is small, with many values close to zero and a few larger observations) and to provide sufficient flexibility of shape, to be used to simulate appropriately distributed residue data for use in the power analysis.

Data sets simulated and analysed within the power analysis mostly involved the comparison of just two samples, but in a few cases either three or five samples were included. For each data set, samples were drawn at random from Gamma distributions with specified values for the mean and variance parameters, with the constraint that within any data set, the variance parameter was kept constant across the distributions.

Assessment of the power of the analysis method was made for a range of sample sizes (20, 40, 60, 80, 100, 150, 200 observations/residues per sample), and for a range of different threshold sets, constructed on a semi-log scale (0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1.0, etc.). The differences between the threshold sets were primarily associated with the lowest threshold used, simulating the limit of determination (LOD), and ranging from 0.5 down to 0.01. An additional constraint was imposed such that the maximum threshold value was no more than five times the value of the mean for the Gamma distribution with the largest mean, i.e., if the mean value was 0.1, then the maximum threshold value used to categorise the observations was 0.5. These constraints affected the number of categories into which each sample of values was divided, indirectly affecting the power of the analysis approach.

For each of the combinations of sample size, threshold set, and sets of Gamma distribution parameters, 1000 separate data sets were simulated and analysed. The power analyses were summarised by considering the proportion of these 1000 data sets which led to a significant test statistic from the Mead-Kift analysis approach. For the comparison of pairs of samples with different mean parameter values, power curves were constructed by plotting the proportions of data sets giving significant test statistics against the difference between the mean parameters for the distributions from which the samples were drawn, thus showing the probabilities of detecting particular sizes of shifts in mean residue levels. Similar summaries were produced for scenarios comparing more than two samples.

Distributions, Thresholds and Samples

As noted above, the choice of thresholds can have a large impact on the power of the method to detect differences – if too many of the categories have zero or very low counts, then this will tend to lead to a reduction in the power of the method. The stacked bar-charts below show the expected proportions of observations from different Gamma distributions that would fall into each of the threshold categories. Figure 9 shows this for Gamma distributions with a range of mean values (0.1 up to 3.2) where the variance has been constrained to be equal to the mean.

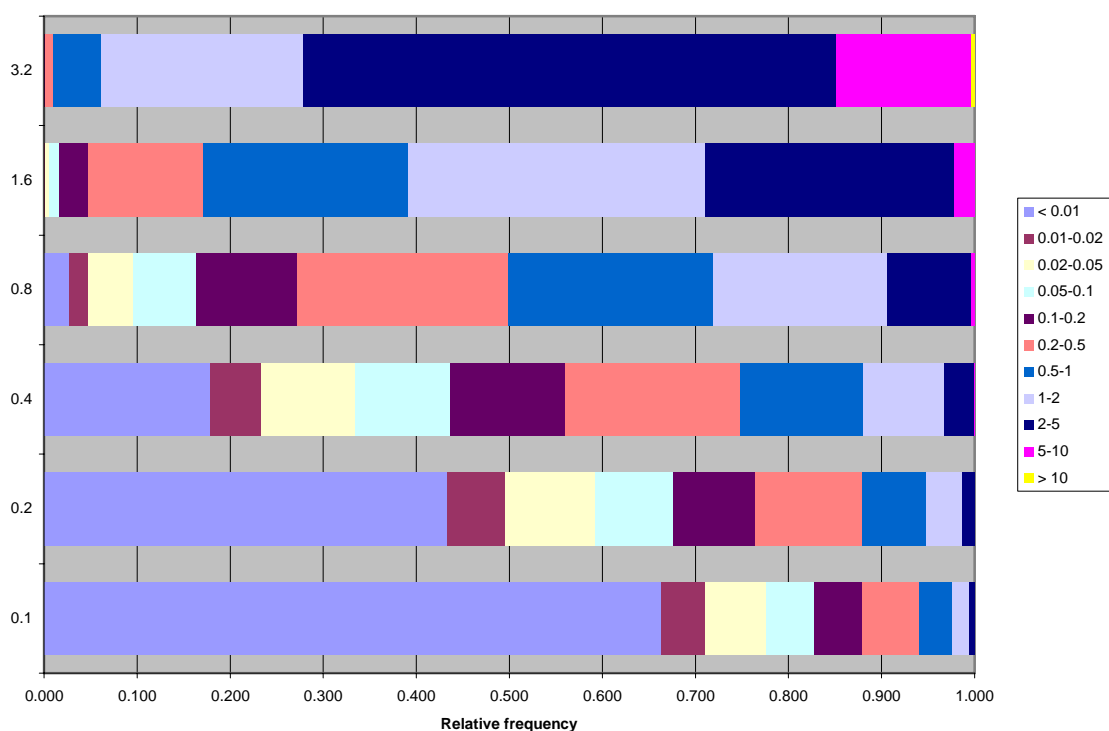


Figure 9: Expected proportions of observations in each of 11 ranges for six different Gamma distributions (different mean values with variance = mean).

Where the mean is small (0.1), the majority of observations will be less than even the lowest threshold (0.01), so that higher categories will have relatively low counts even for quite large sample sizes. As the mean (and variance) increase, the number of categories containing large proportions of observations decreases, with most observations greater than the largest of the LOD values considered (0.5). Thus samples from distributions with relatively large means

(and variances) will generally have a lower power to detect changes using the approach described above, as many of the lower categories (0.01 – 0.02, 0.02 – 0.05, etc.) will have either zero or low counts, thus contributing little to the deviance but something to the degrees of freedom (see “Limitations of approach” above).

Of course sample size will also have an impact here, with larger samples being less likely to result in categories with zero or low counts, even if the proportion of observations in a particular category is expected to be low.

Figure 10 similarly shows the distributions of values across categories for a range of Gamma distributions with different means (0.1 up to 0.6), but with the same variance (= 0.1). As the mean increases, the number of categories containing large proportions of observations again decreases, with the proportion below any particular LOD value (0.01 up to 0.5) decreasing rapidly as the mean increases. Thus we might anticipate that both the multiple threshold and simple threshold analyses might detect changes in the mean residue level here.

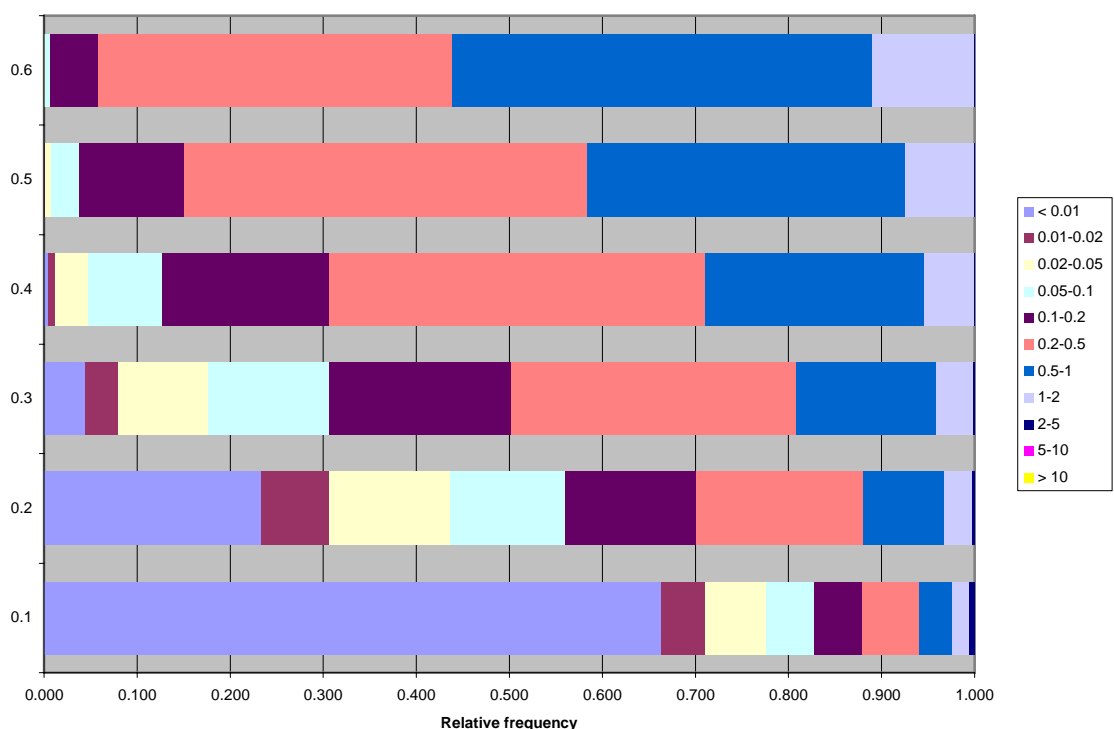


Figure 10: Expected proportions of observations in each of 11 ranges for six different Gamma distributions (different mean values with variance = 0.1).

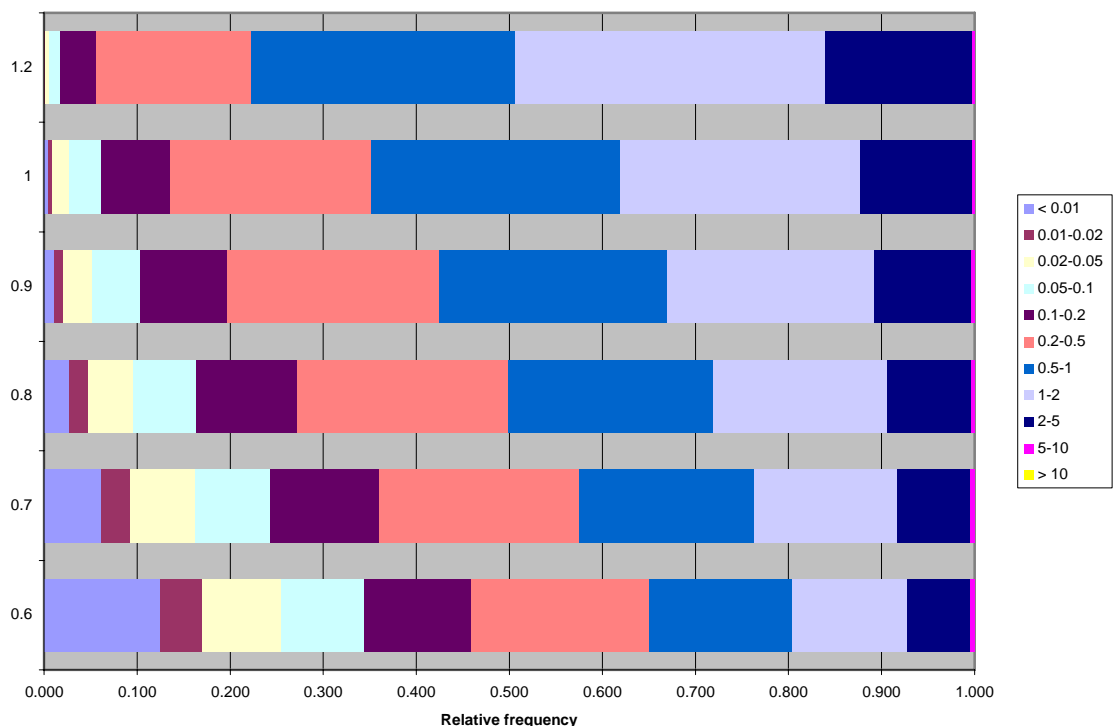


Figure 11: Expected proportions of observations in each of 11 ranges for six different Gamma distributions (different mean values with variance = 0.8).

However for distributions with a larger variance (= 0.8), the proportions below particular LOD thresholds changes less rapidly with changes in the mean residue level (Figure 11), so here we might expect the multiple threshold approach to be more beneficial compared to the single threshold analysis.

Results

Comparison of samples drawn from the same distributions

An important part of the power analysis is to assess how often samples drawn from the same distributions will be identified as being different. If we assess the test result at a 5% significance level (so that the test is defined to have a size of 5%) then we would expect the analysis to indicate a significant difference between samples for 5% of our data sets.

Table 11 shows the proportion of significant test results at a 5% significance level when comparing two samples from Gamma distributions with mean = variance = 0.1, for all combinations of a range of sample sizes (20 up to 200) and for different limits of determination. Note that as the limit of determination decreases, the simulated residues are classified into a larger number of categories.

Table 11: Simulation of two samples drawn from Gamma distributions with equal means, and mean = variance = 0.1. Proportion of comparisons (based on 1000 simulations) giving test results significant at $P < 0.05$ – effects of sample size and simulated limit of determination (LOD).

Limit of determination	0.01	0.02	0.05	0.1	0.2	0.5
Number of categories	7	6	5	4	3	2
Sample size						
20	0.064	0.075	0.062	0.051	0.050	0.062
40	0.088	0.099	0.089	0.076	0.082	0.084
60	0.097	0.073	0.084	0.071	0.050	0.078
80	0.079	0.063	0.062	0.072	0.067	0.056
100	0.082	0.069	0.071	0.077	0.072	0.063
150	0.058	0.050	0.051	0.055	0.051	0.056
200	0.060	0.057	0.053	0.055	0.042	0.045

Certainly for large samples sizes (150, 200) and relatively high values of the limit of determination, the proportions are quite close to the 0.05 value we would expect. For lower limits of determination and smaller sample sizes, the proportion of significant results is somewhat larger than the expected 5%, possibly because of the relatively small counts for some categories (most observations will be below the LOD for all of the LOD thresholds) resulting in large relative differences in the proportions in these categories between samples. Example samples for samples of size 60 with the limit of determination at 0.01 are shown in Table 12.

Table 12: Example samples of size 60 drawn from a Gamma distribution with mean = variance = 0.1

Group	Sample									
	1	2	3	4	5	6	7	8	9	10
<0.01	44	40	43	44	38	29	41	38	38	42
0.01 – 0.02	1	2	1	1	1	3	3	4	3	4
0.02 – 0.05	3	4	6	6	3	6	3	4	6	3
0.05 – 0.1	2	2	3	2	1	8	1	3	4	5
0.1 – 0.2	2	3	2	2	4	1	2	3	3	4
0.2 – 0.5	2	4	0	2	7	10	4	6	2	2
> 0.5	6	5	5	3	6	3	6	2	4	0

Note the greater relative variation in counts in the “0.2 -0.5” and “> 0.5” categories between samples than in the “< 0.01” category.

Similar patterns of proportions of significant results when comparing samples from the same distributions can be seen for Gamma distributions with larger means (= variance = 0.4 or 0.8), though the increase over the expected proportion of significant results is generally lower (Tables 13 and 14).

Table 13: Simulation of two samples drawn from Gamma distributions with equal means, and mean = variance = 0.4. Proportion of comparisons (based on 1000 simulations) giving test results significant at $P < 0.05$ – effects of sample size and simulated limit of determination (LOD)

Limit of determination	0.01	0.02	0.05	0.1	0.2	0.5
Number of categories	7	6	5	4	3	2
Sample size						
20	0.085	0.090	0.091	0.090	0.067	0.064
40	0.077	0.089	0.079	0.067	0.073	0.080
60	0.077	0.067	0.077	0.065	0.080	0.070
80	0.072	0.070	0.072	0.065	0.067	0.071
100	0.065	0.071	0.072	0.072	0.048	0.058
150	0.058	0.056	0.065	0.058	0.045	0.054
200	0.062	0.057	0.058	0.050	0.047	0.050

Table 14: Simulation of two samples from Gamma distributions with equal means, and mean = variance = 0.8. Proportion of comparisons (based on 1000 simulations) giving test results significant at $P < 0.05$ – effects of sample size and simulated limit of determination (LOD)

Limit of determination	0.01	0.02	0.05	0.1	0.2	0.5
Number of categories	7	6	5	4	3	2
Sample size						
20	0.058	0.084	0.079	0.104	0.069	0.068
40	0.075	0.101	0.088	0.063	0.065	0.069
60	0.076	0.090	0.064	0.070	0.054	0.058
80	0.072	0.059	0.055	0.058	0.054	0.049
100	0.078	0.059	0.062	0.047	0.063	0.061
150	0.069	0.076	0.046	0.051	0.056	0.053
200	0.060	0.057	0.052	0.049	0.043	0.049

For a given sample size, as the Gamma distribution mean (= variance) increases, the method tends to result in fewer significant results than would be expected (see Table 15) particularly when the limit of determination is low, probably because both samples have zero or very low counts in a number of categories close to the limit of determination:

Table 15: Simulation of two samples drawn from Gamma distributions with equal means, mean = variance and sample size = 40. Proportion of comparisons (based on 1000 simulations) giving test results significant at $P < 0.05$ – effects of mean (= variance) and simulated limit of determination (LOD). The number of categories used in each analysis is shown in parentheses.

Limit of determination	0.01	0.02	0.05	0.1	0.2	0.5
Mean						
0.1	0.088 (7)	0.099 (6)	0.089 (5)	0.076 (4)	0.082 (3)	0.084 (2)
0.2	0.086 (8)	0.088 (7)	0.082 (6)	0.079 (5)	0.079 (4)	0.073 (3)
0.4	0.077 (9)	0.089 (8)	0.079 (7)	0.067 (6)	0.073 (5)	0.080 (4)
0.8	0.075 (9)	0.101 (8)	0.088 (7)	0.063 (6)	0.065 (5)	0.069 (4)
1.6	0.010 (10)	0.018 (9)	0.030 (8)	0.046 (7)	0.073 (6)	0.057 (5)
3.2	0.001 (11)	0.002 (10)	0.003 (9)	0.008 (8)	0.019 (7)	0.032 (6)

Similar assessments of the analysis method can also be made where more than two samples drawn from the same distribution are being compared. Tables 16 and 17 show a sample of simulation results for the analysis of data sets comprising either three or five samples, respectively, drawn from Gamma distributions with mean = variance, and considering a range of different values for the mean.

Table 16: Simulation of three samples drawn from Gamma distributions with equal means, mean = variance and sample size = 40. Proportion of comparisons (based on 1000 simulations) giving test results significant at $P < 0.05$ – effects of mean (= variance) and simulated limit of determination (LOD). The number of categories used in each analysis is shown in parentheses.

Limit of determination	0.01	0.02	0.05	0.1	0.2	0.5
Mean						
0.1	0.113 (7)	0.101 (6)	0.085 (5)	0.102 (4)	0.083 (3)	0.096 (2)
0.2	0.091 (8)	0.093 (7)	0.097 (6)	0.080 (5)	0.068 (4)	0.087 (3)
0.4	0.084 (9)	0.080 (8)	0.081 (7)	0.093 (6)	0.081 (5)	0.061 (4)
0.8	0.077 (9)	0.098 (8)	0.080 (7)	0.078 (6)	0.068 (5)	0.061 (4)
1.6	0.003 (10)	0.019 (9)	0.032 (8)	0.050 (7)	0.058 (6)	0.052 (5)
3.2	0.000 (11)	0.001 (10)	0.001 (9)	0.006 (8)	0.010 (7)	0.029 (6)

Table 17: Simulation of five samples drawn from Gamma distributions with equal means, mean = variance and sample size = 40. Proportion of comparisons (based on 1000 simulations) giving test results significant at $P < 0.05$ – effects of mean (= variance) and simulated limit of determination (LOD). The number of categories used in each analysis is shown in parentheses.

Limit of determination	0.01	0.02	0.05	0.1	0.2	0.5
Mean						
0.1	0.129 (7)	0.126 (6)	0.119 (5)	0.086 (4)	0.075 (3)	0.063 (2)
0.2	0.112 (8)	0.095 (7)	0.11 (6)	0.090 (5)	0.107 (4)	0.094 (3)
0.4	0.104 (9)	0.090 (8)	0.077 (7)	0.082 (6)	0.081 (5)	0.069 (4)
0.8	0.103 (9)	0.105 (8)	0.092 (7)	0.067 (6)	0.074 (5)	0.055 (4)
1.6	0.003 (10)	0.009 (9)	0.034 (8)	0.071 (7)	0.079 (6)	0.066 (5)
3.2	0.000 (11)	0.000 (10)	0.000 (9)	0.001 (8)	0.005 (7)	0.025 (6)

Again, the analysis method tends to result in a larger proportion of significant results than expected for samples drawn from distributions with low means (0.1 to 0.8) and much smaller proportions of significant results than expected for samples drawn from distributions with high means (1.6, 3.2). Further exploration of the cause of these results is needed, but was not possible within the constraints of this project.

Comparison of two samples drawn from distributions with different means but the same variance

These comparisons provide a guide to the size of shift in the mean residue level that the analysis method would be able to detect. As yet a similar assessment has not been made for the equivalent single LOD threshold analysis approach.

The power of the analysis method depends on the baseline mean residue level, the chosen limit of determination (and hence the number of categories into which the values are grouped), and the sample size. Figures 12 - 14 show a selection of power curves for different LOD values at selected sample sizes (Figures 12 and 13), and for different sample sizes for a selected LOD value (Figure 14). In each case the figures show the power of the analysis method to detect a difference for the comparison of samples drawn from a Gamma distribution for a range of different mean values but a constant variance of 0.1, with a baseline sample drawn from a Gamma distribution with a mean and variance of 0.1:

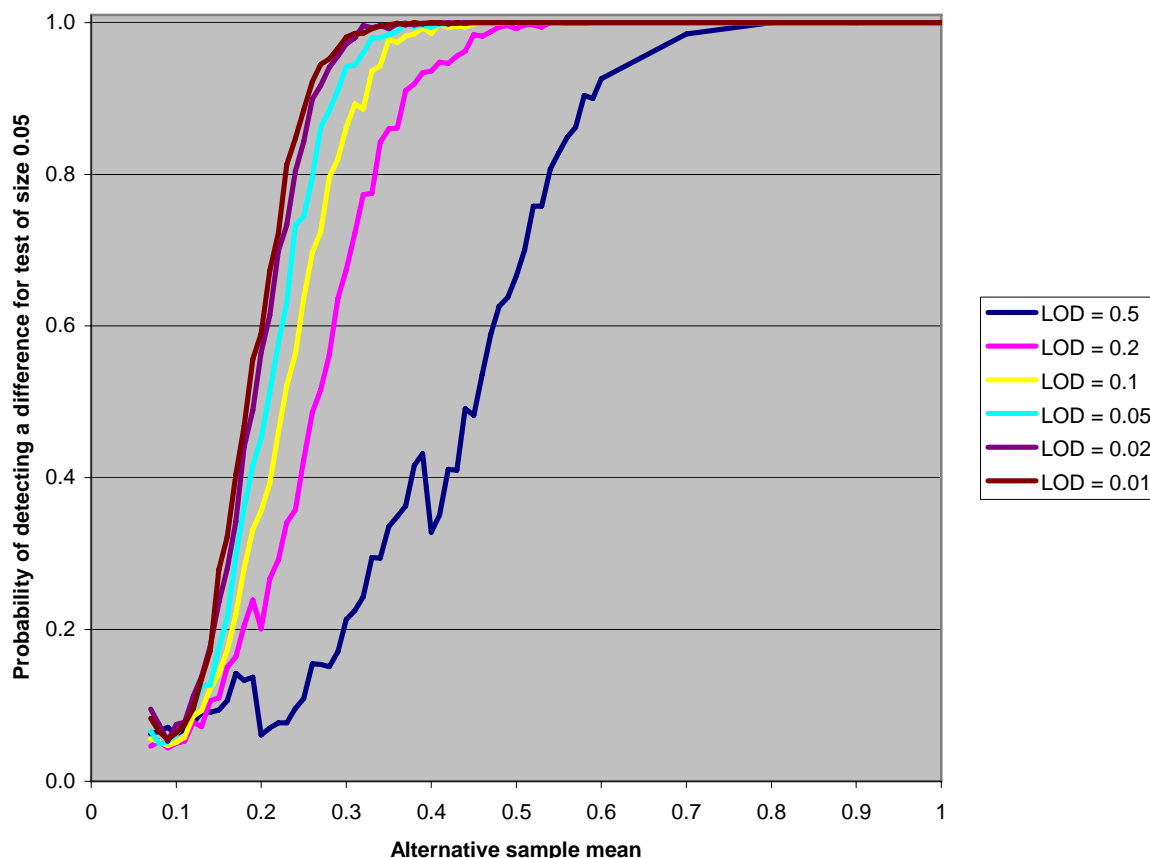


Figure 12: Simulated power curves for the comparisons of two samples of size 20 both drawn from Gamma distributions with variance = 0.1. One sample is drawn from a distribution with mean = 0.1, with the second sample drawn from a distribution with mean as indicated on the horizontal axis. Power curves are presented for analyses with six different levels of determination (LOD).

Note that the non-monotonic nature of the power curve for “LOD = 0.5” is caused by changes in the number of categories on which the analysis was based – all the curves would be much smoother if the number of categories was kept constant across the analyses for all alternative sample means.

Note also that reducing the limit of determination results in a great improvement in power until the LOD is smaller than the mean residue level for the base-line sample, after which little increase in power is seen.

But even with a sample size of only 20 the analysis method is able to detect a change in mean residue from 0.3 to 0.1 if the limit of determination is low enough (0.05 or smaller).

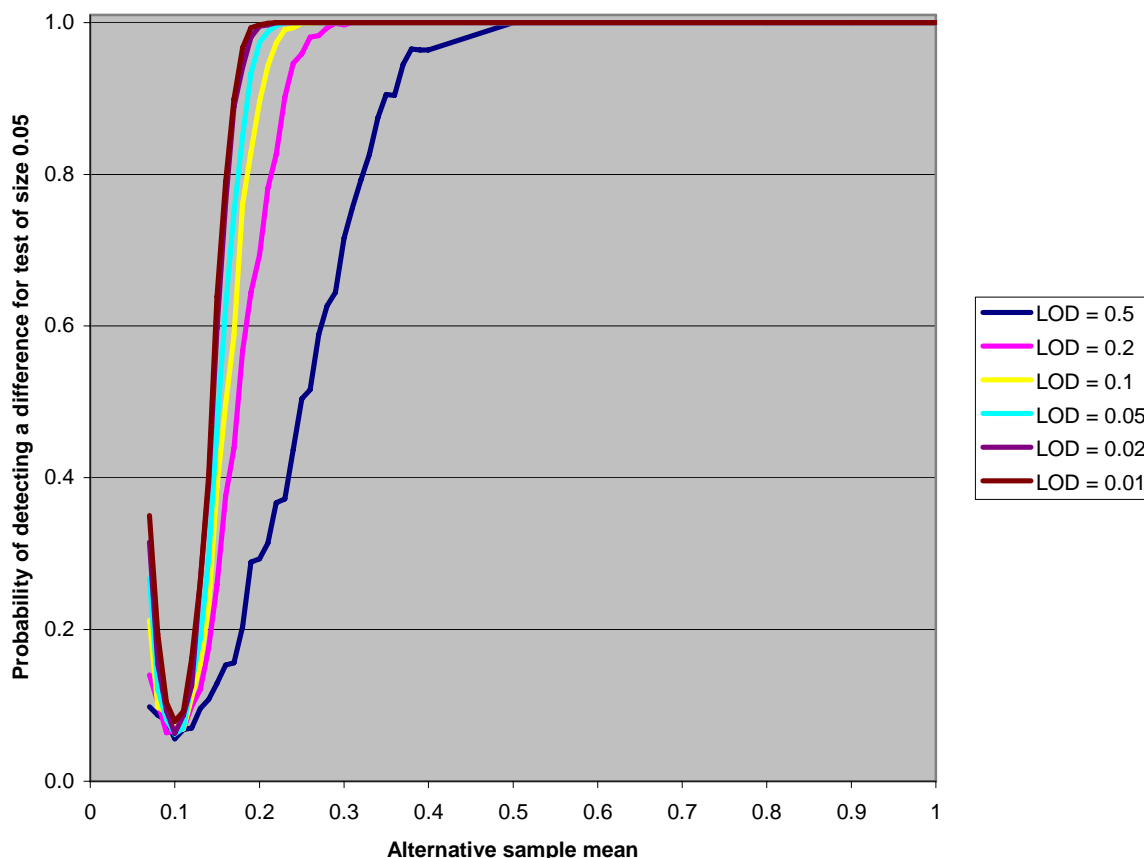


Figure 13: Simulated power curves for the comparisons of two samples of size 80 both drawn from Gamma distributions with variance = 0.1. One sample is drawn from a distribution with mean = 0.1, with the second sample drawn from a distribution with mean as indicated on the horizontal axis. Power curves are presented for analyses with six different levels of determination (LOD).

Increasing the sample size to 80 (Figure 13) further increases the power of the analysis method (an expected result as most tests will have increased power with an increased number of observations), with changes in mean residue level from 0.2 to 0.1 now being readily detectable.

Considering the power curves for a range of sample sizes at an LOD of 0.1 more clearly shows the impact of sample size on the power of the method (Figure 14)

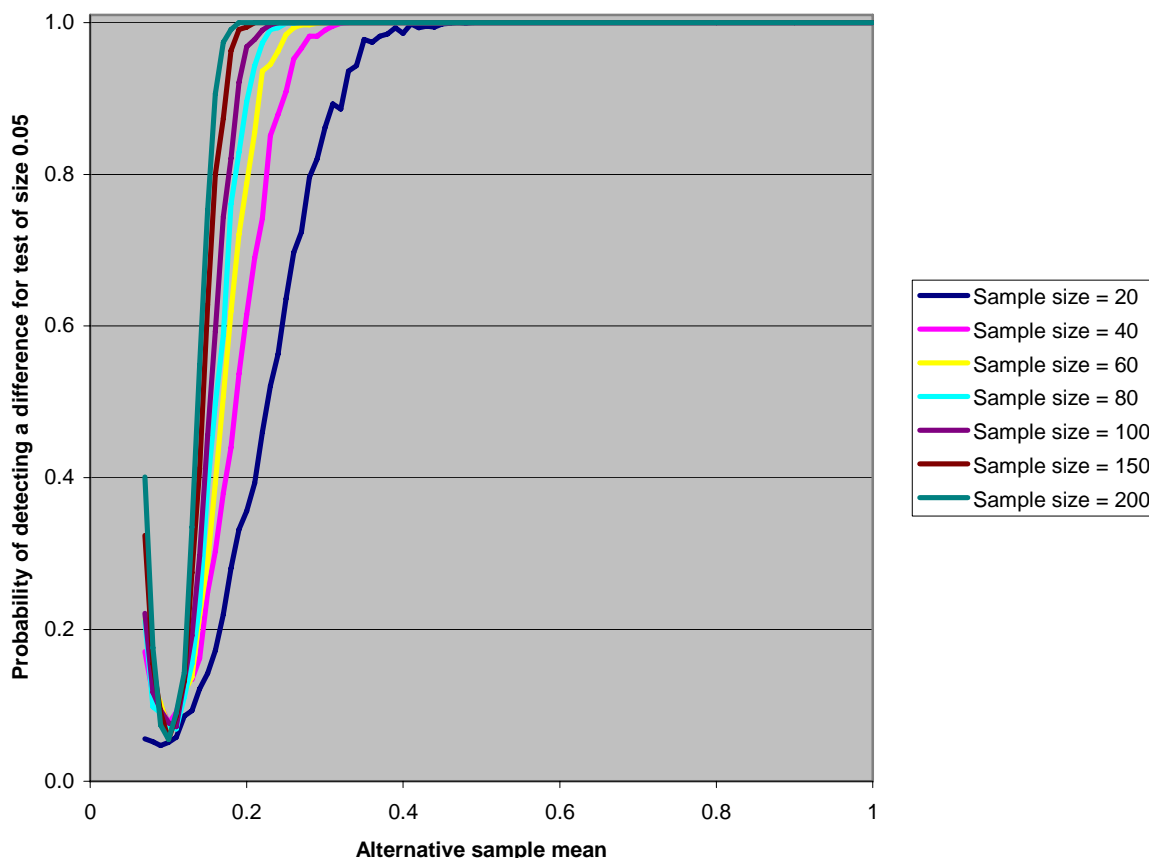


Figure 14: Simulated power curves for the comparisons of two samples of the same size, both drawn from Gamma distributions with variance = 0.1. All analyses were performed with a limit of determination (LOD) of 0.1. One sample is drawn from a distribution with mean = 0.1, with the second sample drawn from a distribution with mean as indicated on the horizontal axis. Power curves are presented for comparisons using six different sample sizes

With a larger base-line mean residue level (= 0.4) and a correspondingly larger Gamma distribution variance for all samples, the effect of varying the limit of determination is less dramatic, but the effect of sample size is still important. A selection of power curves for these scenarios are shown in Figures 15 and 16.

In Figure 15, showing the effects of different LOD values for comparisons using a constant sample size of 20, notice that the power curves for the two lowest LOD values are actually below those for the intermediate LOD values. This reduction in power is almost certainly because the lowest categories have zero counts for both samples, thus increasing the degrees of freedom for the assessment of differences between the samples without increasing the deviance.

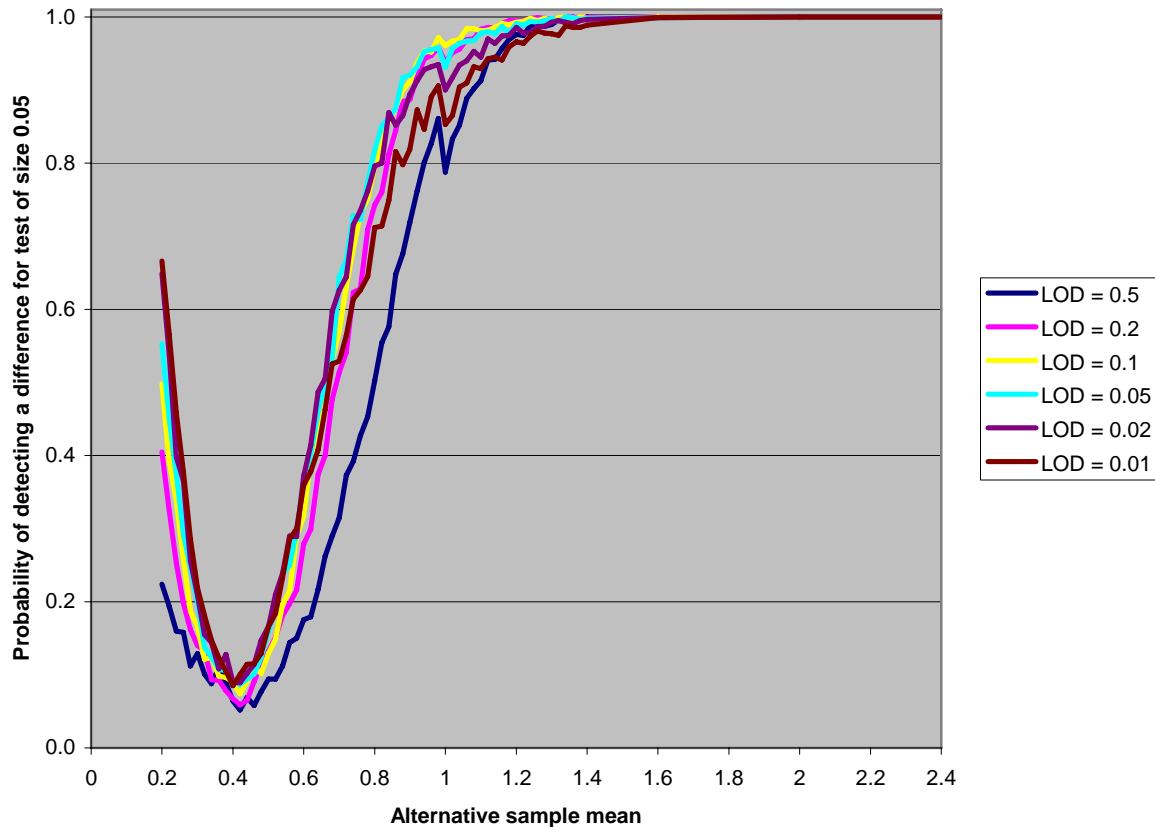


Figure 15: Simulated power curves for the comparisons of two samples of size 20, both drawn from Gamma distributions with variance = 0.4. One sample is drawn from a distribution with mean = 0.4, with the second sample drawn from a distribution with mean as indicated on the horizontal axis. Power curves are presented for analyses with six different levels of determination (LOD).

In Figure 16, showing the effect of different sample sizes for a threshold set based on an LOD of 0.1, notice that the additional benefit of increasing sample size is now relatively small above a sample size of about 100.

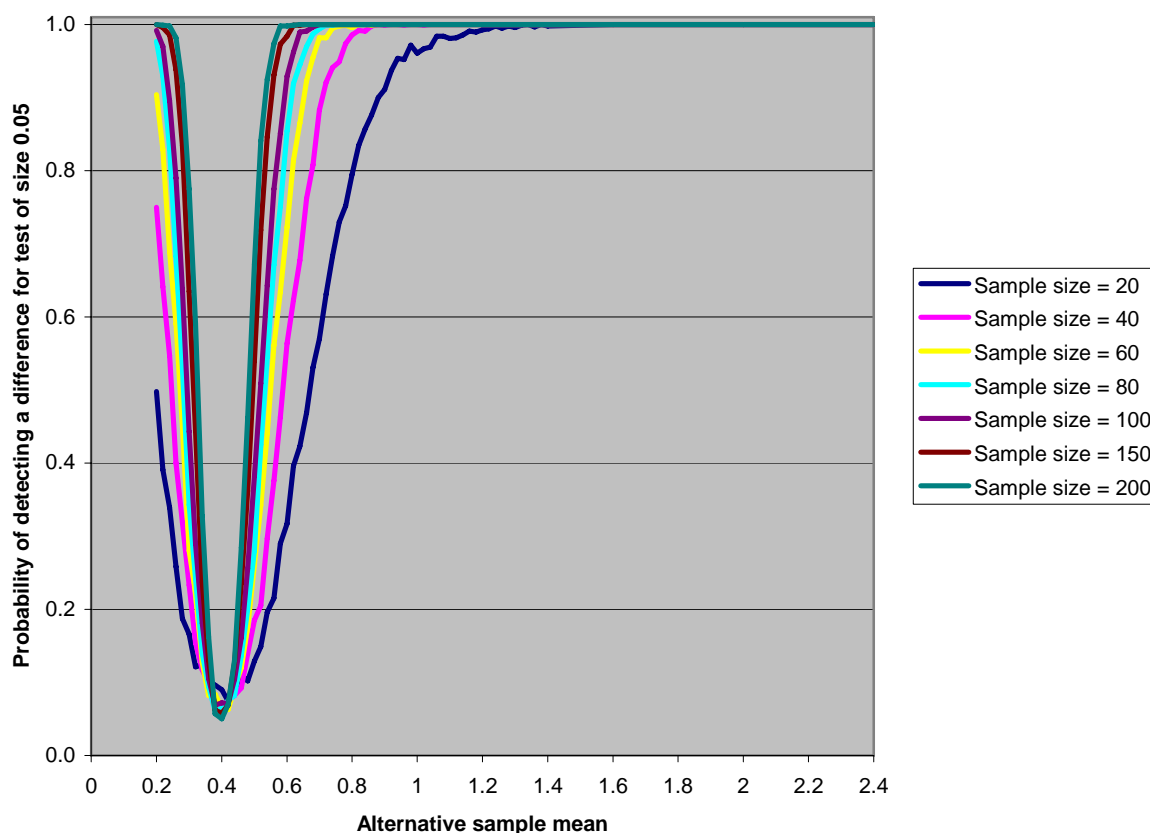


Figure 16: Simulated power curves for the comparisons of two samples of the same size, both drawn from Gamma distributions with variance = 0.4. All analyses were performed with a limit of determination (LOD) of 0.1. One sample is drawn from a distribution with mean = 0.4, with the second sample drawn from a distribution with mean as indicated on the horizontal axis. Power curves are presented for comparisons using six different sample sizes

Further analyses have been performed with larger simulated mean residue values (and larger Gamma distribution variance values), but the results follow similar patterns to those seen above for a base-line mean residue level of 0.4, and so are not presented here.

Comparison of multiple samples drawn from distributions with different means but the same variance

As the analysis method is capable of simultaneously comparing the distributions of residue values from multiple samples, it seemed appropriate to also assess the power of the method to detect differences between more than two samples. However, a simple graphical presentation is then no longer possible, certainly if there are more than three samples.

Tables 18, 19 and 20 show the proportion of significant results (the power of the analysis method) when comparing the distributions of three samples drawn from Gamma distributions

with equal variances. The first sample is always drawn from a distribution with the mean equal to the variance, and the means for the other two samples are as indicated in the tables.

Table 18 shows the results for comparisons where the first, base-line, sample is drawn from a Gamma distribution with mean = variance = 0.1, with all three samples of size 20, with an intermediate LOD value of 0.1

Table 18: Simulation of three samples drawn from Gamma distributions with variance = 0.1 and sample size = 20. First sample drawn from a distribution with mean = 0.1, with the other samples drawn from distributions with means as specified in the table. All analyses performed with LOD = 0.1. Proportion of comparisons (based on 1000 simulations) giving test results significant at $P < 0.05$ – effects of varying means for second and third samples.

	Sample mean 3					
Sample mean 2	0.1	0.2	0.3	0.4	0.6	1.0
0.1	0.059					
0.2		0.323				
0.3		0.762	0.880			
0.4		0.955	0.968	0.995		
0.6		1.000	1.000	1.000	1.000	
1.0		1.000	1.000	1.000	1.000	1.000

Notice that if all three samples are from the same distribution then a slightly higher than expected proportion of tests are significant, but that the test is able to detect quite small differences if both the other two samples are from distributions with larger means. Further, more detailed studies (smaller changes in mean, wide range of combinations of means) are required for this scenario to fully explore the power of the method.

With a larger baseline mean and Gamma distribution variance, the power is smaller for the same sample size of 20 (as might be expected from the two-sample power analyses above) (Table 19), but a similarly powerful test can be achieved with an increase in sample size (Table 20).

Table 19: Simulation of three samples drawn from Gamma distributions with variance = 0.4 and sample size = 20. First sample drawn from a distribution with mean = 0.4, with the other samples drawn from distributions with means as specified in the table. All analyses performed with LOD = 0.1. Proportion of comparisons (based on 1000 simulations) giving test results significant at $P < 0.05$ – effects of varying means for second and third samples.

	Sample mean 3					
Sample mean 2	0.4	0.5	0.6	0.8	1.0	1.4
0.4	0.076					
0.5		0.139				
0.6		0.254	0.352			
0.8		0.709	0.685	0.852		
1.0		0.904	0.889	0.925	0.968	
1.4		1.000	0.997	0.994	1.000	1.000

Table 20: Simulation of three samples drawn from Gamma distributions with variance = 0.4 and sample size = 60. First sample drawn from a distribution with mean = 0.4, with the other samples drawn from distributions with means as specified in the table. All analyses performed with LOD = 0.1. Proportion of comparisons (based on 1000 simulations) giving test results significant at $P < 0.05$ – effects of varying means for second and third samples.

	Sample mean 3					
Sample mean 2	0.4	0.5	0.6	0.8	1.0	1.4
0.4	0.067					
0.5		0.246				
0.6		0.630	0.752			
0.8		0.999	0.996	1.000		
1.0		1.000	1.000	1.000	1.000	
1.4		1.000	1.000	1.000	1.000	1.000

Again, further, more detailed studies are needed to fully understand the power of the method to detect differences between multiple samples, ideally incorporating some sort of fixed trend between samples, based on expectation of how residue levels might be changed by pesticide usage practices.

But, the conclusions from the power analysis must be that if the Gamma distribution does indeed provide a good fit to most sets of pesticide residue data, then the analysis method should be able to detect relatively small changes in the mean pesticide residue level between samples using sample sizes ranging from 20 – 100 and with limits of determination that are ideally slightly lower than the minimum mean residue level achieved.

References

Kift, N.B., Mead, A., Reynolds, K., Sime, S., Barber, M.D., Denholm, I & Tatchell, G.M. (2004). The impact of insecticide resistance in the currant-lettuce aphid, *Nasonovia ribisnigri*, on pest management in lettuce. *Agricultural and Forest Entomology*, **6**, 295-309.