Discrimination of Gyrodactylids Based on Landmark Data

J.W. Kay[1] and A.P. Shinn [2]
[1] Department of Statistics, University of Glasgow, G12 8QQ, UK
[2] Institute of Aquaculture, University of Stirling, FK9 4LA, UK

1 Introduction

Gyrodactylids are parasites which attach themselves to the skin and fins of fresh water fish. A scanning-electron microscope image of a gyrodactylid is given in Fig.1 (left). Each gyrodactylid has an attachment organ by means of which it anchors itself onto its host. This attachment organ contains three distinct sclerite structures, namely, the hamuli , the ventral bars and the marginal hooks as illustrated in Fig. 1 (right).

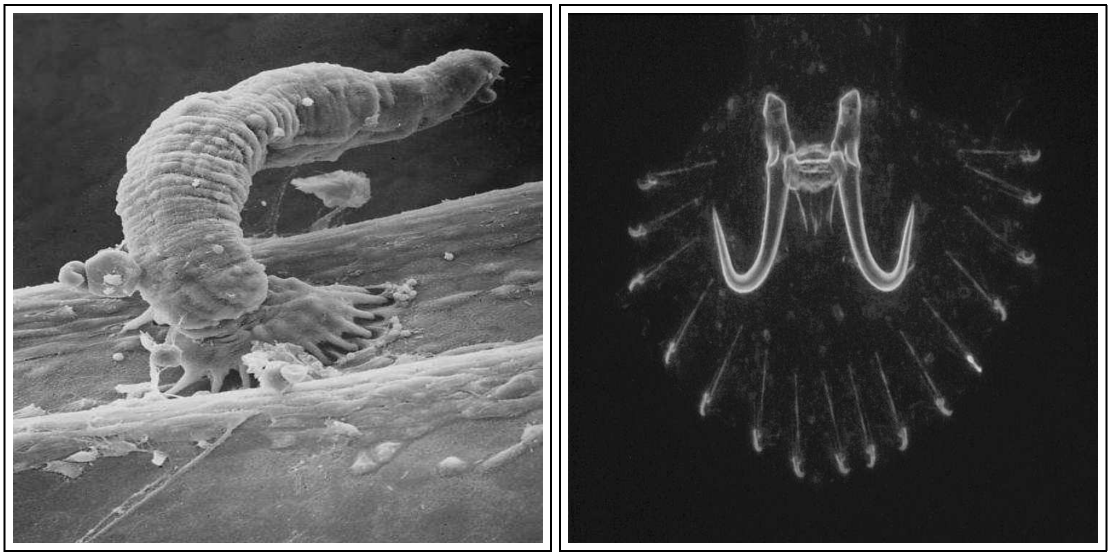

Figure 1: Left: a gyrodactylid attached to a fin. Right: a phase-contrast image of the attachment organ of *Gyrodactylus*. The central complex comprises two large hamuli linked by two connecting bars, the dorsal and ventral bars, but the principal force of attachment is realised by the sixteen peripherally-positioned marginal hooks.

There are many different species of the genus *Gyrodactylus*. One particular form, *Gyrodactylus salaris*, is known to be highly pathogenic to stocks of Atlantic salmon, whereas other species that infect salmonids have a generally low pathogenicity. *Gyrodactylus salaris* is responsible for the catastrophic decline in salmon stocks in Norway and has been demonstrated to be widespread in Norwegian rivers. It has also caused problems in Portugal and France. In order to prevent its entry into the UK, G. *salaris* was made a notifiable disease in 1988 under the 1937 and 1983 Diseases of Fish Acts of the UK. While the UK is thought to be free of G. *salaris* there is another species, G. *thymalli* which has been found in the UK and some think is a variant of G. *salaris*. It is important to find a means of identification of G. *salaris* via routine microscopic monitoring of samples of parasites. Hence the main motivation for this work is the development of a statistical method which could be used to discriminate G. *salaris* from other species of *Gyrodactylus*, while a secondary aim is the discrimination of the other species of *Gyrodactylus* from each other.

In earlier work, morphometric measurements were painstakingly recorded from each of the three sclerite structures of samples of specimens of *Gyrodactylus* and recent work (Kay et al., 1999, Shinn et al., 2000, McHugh et al., 2000) has shown that the application of nonlinear statistical classifiers with data from such morphometric measurements have promising potential for discriminating G. *salaris* from other other species of *Gyrodactylus* and also, to some extent, distinguishing some other species of *Gyrodactylus* from each other. Current work on this problem centres on the development of a microscopy-based semi-automatic system which could be used routinely for species identification. This work will involve various techniques including image analysis, object recognition and statistical size-and-shape analysis. The work described herein makes use of landmark data of the hamulus which have been extracted from images obtained by bright-field microscopy. Two main approaches to the discrimination of seven different species of *Gyrodactylus* that are built on standard methods (see, for example, McLachlan, 1992) will be discussed. The first is based on the EDMA approach to statistical shape analysis approach and the second is an adaptation of the standard k nearest neighbours algorithm in which reflection size-and-shape Procrustes distance is used to measure distance between pairs of landmark configurations. See Bookstein (1991), Dryden and Mardia (1998) and Lele and Richtsmeier (2001) for the background on statistical shape analysis.

2 The Data

In this initial study we consider landmark data obtained from the hamuli of a set of 88 specimens, each of which is known (by expert opinion) to belong to one and only one of seven species. This set contains 20 specimens of G. *salaris*, 20 of G. *thymalli*, 10 each of G. *colemanensis*, G. *derjavini*, G. *gasterostei* and G. *truttae* and 8 specimens of G. *arcuatus*. Some examples of the hamuli of some of the species are given in Fig. 2. It is clear that these specimens differ in size and shape with the G. *thymalli* and G. *salaris* hamuli being larger than those of the other two hamuli.

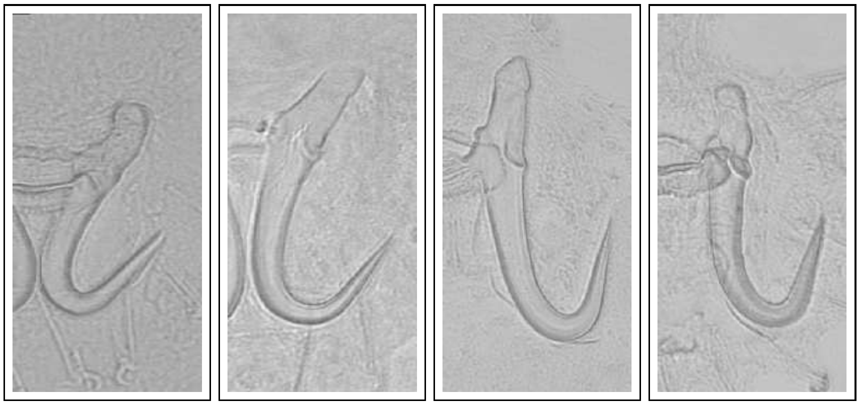

Figure 2: Light microscope images of the hamulus from four species of *Gyrodactylus*. From left to right: G. *derjavini*, G. *salaris*, G. *thymalli* and G. *truttae*.

The single hamuli in the images considered in this study can be presented at different translations, rotations as well as being reflected, as illustrated in Fig. 3. Hence any data analysis performed on extracted landmark co-ordinates must be invariant to reflections, rotations and translations and thus would constitute a reflection size-and-shape analysis (Dryden and Mardia, 1998; p. 57).

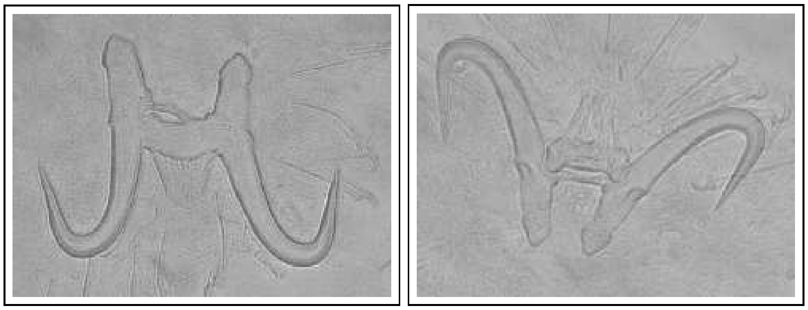

Figure 3: Light microscope images of hamuli of G. *thymalli* specimens in different orientations with different reflections of individual hamuli.

Six landmarks have been identified on the hamulus and these are defined in Fig. 4. So for each specimen we have available a 6 × 2 matrix of landmark co-ordinates, termed a landmark configuration. The co-ordinates were extracted manually from light microscope images of the specimens.

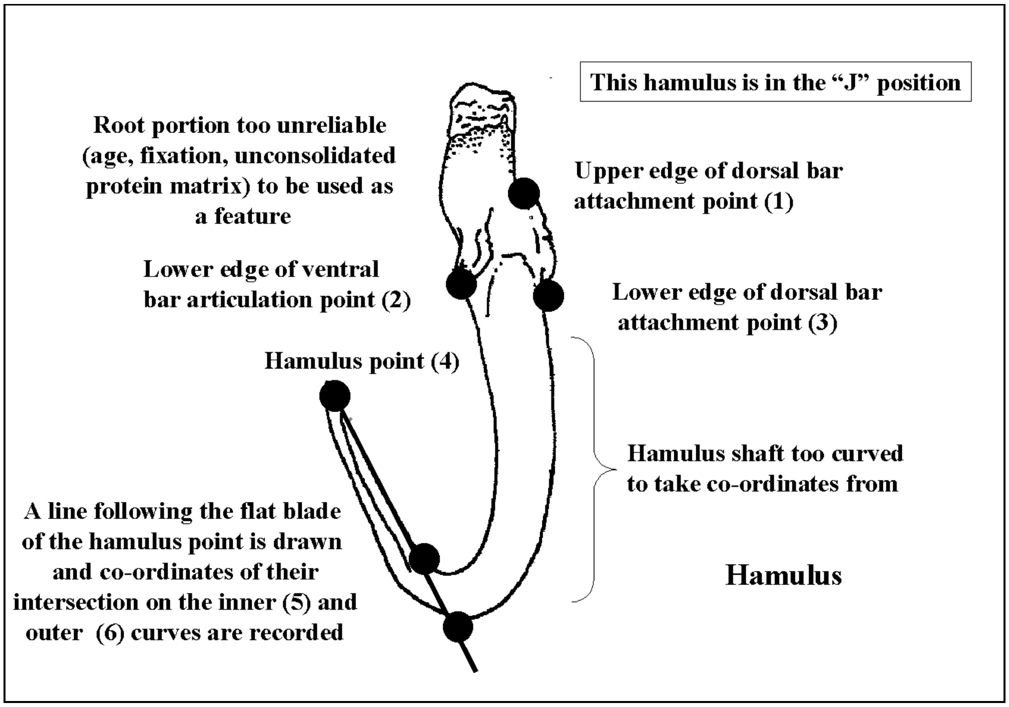
This hamulus is in the "J" position
Root portion too unreliable (age, fixation, unconsolidated protein matrix) to be used as a feature
Upper edge of dorsal bar attachment point (1)
Lower edge of ventral bar articulation point (2)
Lower edge of dorsal bar attachment point (3)
Hamulus point (4)
Hamulus shaft too curved to take co-ordinates from
A line following the flat blade of the hamulus point is drawn and co-ordinates of their intersection on the inner (5) and outer (6) curves are recorded
Hamulus

Figure 4: A schematic diagram of a hamulus in J position indicating the positions of the six landmarks.

It is necessary to consider size as well as shape in this discrimination problem. This is illustrated in Fig. 5 which shows that G. *salaris* and G. *thymalli* tend to be larger in terms of centroid size (Dryden and Mardia, 1998; p. 24) than the other species but similar to each other, on average, while G. *arcuatus* is smaller than the other species.

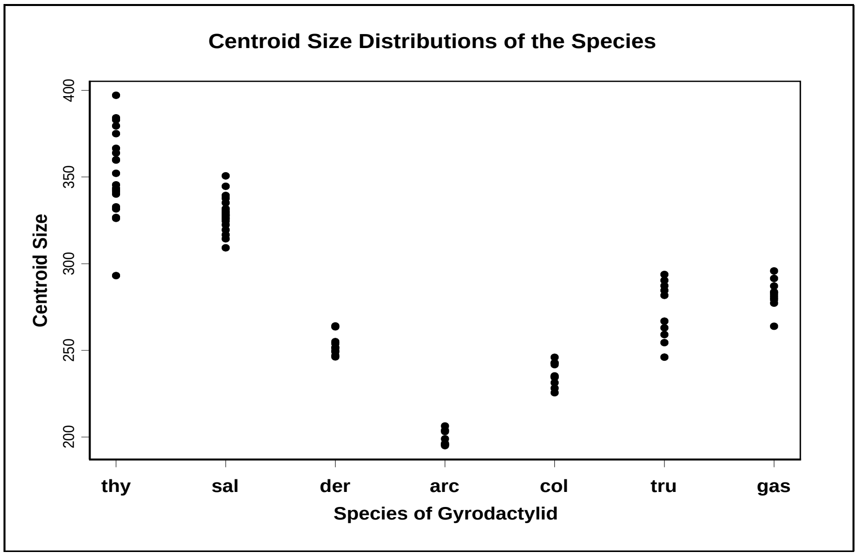

Figure 5: Centroid size distributions of the seven different species.

3 Discrimination using Inter-Landmark Distances

The first approach to discrimination is based on the Euclidean Distance Matrix Analysis (EDMA) approach to statistical shape analysis (Lele and Richtsmeier, 2001). For each landmark configuration, the Euclidean distance between each pair of the k landmarks is computed and these may be presented as a (symmetric) $k \times k$ distance matrix. There are $\frac{1}{2}k(k-1)$ distances and they are invariant with respect to translations, reflections and rotations of the underlying hamuli (as required). Lele and Richtsmeier (2001) do provide a method for the classification of objects based on landmark data; this is essentially a closest-mean classifier in which each landmark configuration is classified as belonging to that class whose mean form it is closest to in terms of shape or size-and-shape.

Here we consider an alternative approach and view the inter-landmark distances obtained for each specimen as the data to which statistical classifiers will be applied. In this application there are 15 inter-landmark distances and so a 15-dimensional ’observation’ vector is available for each specimen, along with its true class identifier. As one might expect, the inter-landmark distances are highly correlated and so not all of the distances are necessary in the discriminant analysis. Indeed, performing a canonical variate analysis results in the first two eigenvalues accounting for 96.3% of the possible linear discriminability. Hence the plot of the inter-landmark distances with respect to the first two canonical variates provides a very good representation of the linear separation among the classes. In Fig. 6 we see that the G. *arcuatus* and G. *colemanensis* specimens are well separated from the other species. The G. *derjavini* and G. *truttae* specimens are quite separate from the remaining species, but quite close to each other. The G. *salaris* separate out from those of G. *truttae* and G. *gasterostei* but are close to these groups.

Forward stepwise linear discriminant analysis was applied using the fifteen inter-landmark distances as the potential discriminating variables. Only six of the distances were used in the final linear classifier, namely the distances between landmarks 1&2, 1&5, 2&5, 2&6, 4&5 and

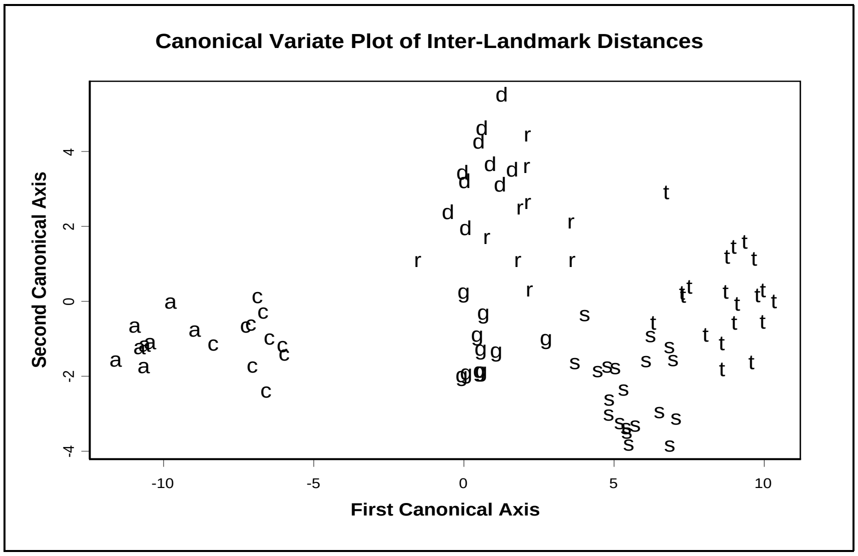

Figure 6: Projection of the inter-landmark distances onto the first two canonical axes, with specimens labelled by the first letter of their species name, apart from G. *truttae* which was labelled as r.

4&6, and so landmark number 3 was not used. When all 88 specimens were used as the training data only seven misclassifications resulted. Using leave-one-out cross-validation resulted in 11 errors out of 88 and an overall estimate of 12.5% as the likely generalisation error were this linear method to be applied to new, similar specimens.

The data were also analysed using a combination canonical variates and k nearest neighbours. Using the scores on the first two canonical variates as input data, and using 5 nearest neighbours, resulted in 4 misclassifications out of 88 for the training data; the leave-one-out cross-validation estimate of generalisation error was 7/88 = 7.95%. In order to obtain more realistic estimates of the likely classification/misclassification rates to be expected with similar new specimens (based admittedly on small sample sizes) stratified three-fold cross-validation was applied, using this combined method. The results were averaged over 100 random samples and are given in Table 1. The overall estimate of misclassification rate is 9.7%. The contribution to this estimate of misclassifications involving G. *salaris* is 2.0%.

	thy	sal	der	arc	col	tru	gas
thy	95.8	4.5	0	0	0	0	0
sal	4.2	95.2	0	0	0	0.1	0
der	0	0	91.2	0	0	29.8	0
arc	0	0	0	98.8	0	0	0.1
col	0	0	0	1.2	100	0	0
tru	0	0.3	8.8	0	0	59.8	4.1
gas	0	0	0	0	0	10.3	89.8

Table 1: Estimated confusion matrix obtained by stratified three-fold cross-validation averaged over 100 random samples using a combination of canonical variates and knn. The entry in the ith row and jth column is the mean percentage of specimens of the species in the jth column that are classified as belonging to the species in the ith row.

Estimates of correct classification are high for G. *thymalli*, G. *salaris*, G. *arcuatus* and G. *colemanensis*, not so high for G. *derjavini* and G. *gastersotei* and poor for G. *truttae*. Apart from some likely confusion with G. *thymalli*, and very slight confusion with G. *truttae*, G. *salaris* is well separated from the other species. There is fairly large amount of confusion between G. *derjavini* and G. *truttae* and between G. *truttae* and G. *gasterostei*.

4 Discrimination using Nearest Neighbours in Procrustes Distance

In this second approach to the discrimination we develop a version of the standard k nearest neighbours algorithm (see, for example, Devivjer and Kittler, 1982) in which the neighbourhoods are defined in terms of a Procrustes reflection size-and-shape distance; clearly any other shape or size-and-shape measure of distance could be employed, as required. A clear account of Procrustes methods is given by Dryden and Mardia (1998). The Procrustes distance used here is invariant with respect to reflections, rotations and translations of the specimens and is defined as follows (Mardia et al., 1979, p. 416).

Let X and Y denote two $k \times m$ landmark configurations and let X_c and Y_c be zero-centered versions of them, obtained by pre-multiplication by a suitable centering matrix such as the Helmert sub-matrix (Dryden and Mardia, 1998, p. 34). Suppose that the singular-value decomposition of the matrix $Y_c^T X_c$ is given by VDU^T. Then the reflection size-and-shape Procrustes distance between the landmark configurations X and Y is defined by

$$\mathrm{rsspd}(X, Y) = \mathrm{tr}(X_c X_c^T) + \mathrm{tr}(Y_c Y_c^T) - 2\mathrm{tr}(D).$$

While the Procrustes rotation that takes X towards Y is not the same as the rotation which takes Y towards X, the distance $\mathrm{rsspd}(X, Y)$ is symmetric in X and Y. We now describe a k nearest neighbours algorithm in which distances between the specimens are obtained via the reflection size-and-shape Procrustes distance.

Suppose that we have available two sets of specimens – a *training* set and a *test* set – and a landmark configuration for each specimen. The algorithm consists of the following steps. For each landmark configuration in the *test* set, the reflection size-and-shape Procrustes distances between the test specimen and each of the specimens in the training set are computed. Then the k members of the training set that are closest to the test specimen in reflection size-and-shape space are determined. Then the classes of these k nearest neighbours are found and the the test specimen is allocated to that class which occurs most frequently among the classes of the k nearest neighbours; ties are broken at random if two or more classes have the maximum number of votes. The algorithm was applied to all 88 specimens using first nearest neighbours. This resulted in 11 misclassifications. Using the leave-one-out cross-validation option also resulted in 11 errors and an estimate of the likely generalisation error of 12.5%. More realistic estimates of likely classification/misclassification error rates were obtain using stratified three-fold cross-validation. The results given in Table 2 are based on means taken over 100 random samples. The overall estimate of misclassification rate is 15.9%. The contribution to this estimate of misclassifications involving G.*salaris* is 4.3%. The pattern of results are similar to those of Table 1 but the estimates suggest greater confusion between the pairs of classes that were confused in Table 1. Clearly the results obtained using this method of discrimination are less good that those obtained in Section 3.

	thy	sal	der	arc	col	tru	gas
thy	86.8	10.8	0	0	0	0	0
sal	8.2	89.3	0	0	0	0.1	0
der	0	0	60.9	0	0	21.6	3.4
arc	0	0	0	100	0	0	0
col	0	0	0	0	100	0	0
tru	4.1	0	39.0	0	0	54.6	11.3
gas	1.0	0	0.1	0	0	23.8	85.3

Table 2: Estimated confusion matrix obtained by stratified three-fold cross-validation averaged over 100 random samples with the Procrustes knn method. The entry in the ith row and jth column is the mean percentage of specimens of the species in the jth column that are classified as belonging to the species in the ith row.

The algorithms used in the analyses reported in Sections 3 & 4 were coded in S-Plus and use was made of the *knn*, *lda* and *predict.lda* functions provided by Venables and Ripley (1997) as well as the *defh* and *centroid.size* functions provided in Ian Dryden's R/S-Plus routines.

5 Conclusions

The results of these initial experiments using landmark data from small sample sizes of seven species of *Gyrodactylus* are quite promising. The discrimination method considered in Section 3 gave better overall results than the Procrustes knn approach, in terms of estimated misclassification rates, with the rate involving G. *salaris* being one-half of that obtained using Procrustes-based knn. Given new, similar specimens it seems that G. *salaris* could be identified with a small chance of error, with the main risk of confusion being between G. *salaris* and G. *thymalli*. Clearly there is quite serious confusion between the pairs G.*derjavini* & G.*truttae* and G.*truttae* & G.*gasterostei*, and the discrimination of the individual species within these pairs of species would not be very reliable. Clearly the cross-validation results and these conclusions are based on small sample sizes and it is necessary to repeat this work with larger, more representative sets of specimens. In addition, landmark data are being collected from light microscope images of the marginal hooks of a variety of specimens; here there are twelve landmarks and so better results might be possible. Work will also be pursued on outline data, using for example Fourier and Wavelet descriptors, and the more-detailed nature of such data may be necessary to discriminate between the most confused species.

References

Bookstein, F.L. (1991). *Morphometric tools for landmark data*. Cambridge University Press. Cambridge.

Devivjer P.A. and Kittler, J. (1982) *Pattern recognition: a statistical approach.* Prentice-Hall. London.

Dryden, I.L. and Mardia, K.V. (1998). *Statistical shape analysis*. Wiley. New York.

Kay, J.W., Shinn, A.P. and Sommerville, C. (1999). Towards an Automated System for the Identification of Notifiable Pathogens: Using *Gyrodactylus salaris* as an example. *Parasitology Today*, **15**, 201-206.

Lele, S.R. and Richtsmeier, J.T. (2001) *An invariant approach to statistical analysis of shapes*. Chapman and Hall/CRC. London.

McHugh, E.S., Shinn, A.P. and Kay, J.W. (2000). Discrimination of the notifiable pathogen *Gyrodactylus salaris* from G. *thymalli* (Monogenea) using statistical classifiers applied to morphometric data. *Parasitology*, **121**, 315-323.

McLachlan, G.J. (1992) *Discriminant analysis and statistical pattern recognition*. Wiley. New York.

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979) *Multivariate analysis*. Academic Press. London.

Shinn, A.P., Kay, J.W. and Sommerville, C. (2000). The use of statistical classifiers for the discrimination of species of the genus *Gyrodactylus* (Monogenea) parasitizing salmonids. *Parasitology*, **120**, 261-269.

Venables, W.N. and Ripley, B.D. (1997) *Modern applied statistics with s-plus (2nd edn)*. Springer-Verlag. New York.