

IN THIS ISSUE

A CLOSE LOOK AT THE HAT MATRIX

The hat matrix is the matrix which permits to compute the fitted values $\hat{y}$ from the observed values y in a regression problem when using the least squares method. This is also the matrix which allows to detect the leverage points by looking at its diagonal entries as proposed by Hoaglin and Welsch in 1978. In the first article of this issue, Annick Clerc-Bérod and Stefan Morgenthaler re-examine the hat matrix and propose a way to identify not only leverage points but also leverage groups.

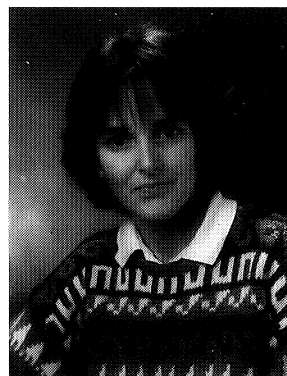


Annick Clerc Bérod was born on May 27, 1964 in Port, Wallis, Switzerland. She obtained her Bachelor of Science in Mathematics from the Swiss Federal Institute of Technology in Lausanne in 1989 and her Ph.D. from the same institute in 1993. She currently works as biostatistician at the Department of Social and Preventive Medicine of the University of Lausanne.

MARKOV CHAINS AND VEGETATION MONITORING

In the second article of this issue Helene Wagner and Otto Wildi use Markov chains to study the development of Swiss dry meadows in terms of species abundance. The statistical problem is to detect a trend in a short multivariate time-series with a large amount of fluctuation. It turns out that Markovian models are a powerful tool to distinguish between trend and noise, and thus to assess a trend in vegetation change.

Helene H. Wagner was born on January 15, 1968 in Zurich. She obtained her Master of Science in Geography from the University of Zurich in 1994. She received her postgraduate degree in Statistics in 1996 from the University of Neuchâtel. Helene Wagner is currently a research assistant at the Swiss Federal Institute for Forest, Snow and Landscape Research in Birmensdorf, Switzerland.



CONTOUR LINES OF L_1 REGRESSION

In the third article of this issue, Jean-Pierre Renfer studies the objective function corresponding to the L_1 regression problem by constructing its contour lines in the parameter space. A simple algorithm has been derived to draw these contour lines.

Jean-Pierre Renfer was born on August 9, 1964 in Neuchâtel. He obtained his Bachelor of Science in Mathematics in 1989 and his Master degree in Statistics in 1990 from the University of Neuchâtel. Jean-Pierre Renfer is currently a research worker in the Statistics Group of the University of Neuchâtel and is teaching applied mathematics to economists.



FROM THE EDITOR'S NOTEBOOK

From the Editor's Notebook we learn how to illustrate graphically the coefficient of determination R^2 in a simple linear regression problem.

IRVING FISHER THE FIRST MATHEMATICAL ECONOMIST OF AMERICA

The semicentennial year of death of Irving Fisher is taking place this year. In *Capturing History*, an homage is paid to this great economist who did use statistics to try to build economics into a genuine science.

CHIKIO HAYASHI AND DATA SCIENCE

The Japanese Statistical Society was established in 1931. The members were almost all economists. Since 1946, many mathematical statisticians joined. Nowadays the composition of members is well-balanced. Most statisticians outside Japan are familiar with names such as Taguchi and Akaike but not much in general know about statistics in Japan and the way they think about statistics. In this issue we are pleased to present to the readers of *Student* the point of view of one of the famous Japanese statisticians Chikio Hayashi on *What is Data Science?*

DATA & STATISTICS

N. Nakamura, N. Ohsumi, K. Onoyama, K. Ogata and R. Taylor publish in this issue a set of measurements obtained from 252 individuals belonging to 82 species of Australian ants called *Myrmecia*. A new data set is also introduced by J. Mexia, L. Gusmão and J. Baeta on "Relative Behavior of a Cohort of Triticale Cultivars".

A Close Look at the Hat Matrix

Annick Clerc Béro

*Social and Preventive Medicine Institute, University of Lausanne,
Switzerland*

Stephan Morgenthaler

Department of Mathematics, EPFL, Switzerland

Abstract: Hoaglin and Welsch (1978) discussed the use of the hat matrix H for the purpose of finding leverage points in regression and ANOVA. A leverage point is any observation with a large diagonal element of H . The present paper extends this work, by exploring a more detailed analysis of H that makes use of the off-diagonal elements as well. A natural measure of outlyingness for a group of design points is the largest eigenvalue of the submatrix of H obtained by retaining the rows and columns corresponding to the indices of the group of interest. This paper discusses a computational strategy for identifying such leverage groups. Our method allows the identification of leverage subsets that would be missed if we only look at the diagonal entries of H .

Key words: Hat matrix, leverage points, leverage subsets.

1 Introduction

We consider the usual linear model $y = X\beta + \epsilon$, where $y_{n \times 1}$ is the n -dimensional vector of the observations, $X_{n \times p}$ is the design matrix, $\beta_{p \times 1}$ is the p -dimensional vector of the unknown parameters, and $\epsilon_{n \times 1}$ the vector of errors. We assume that the errors have zero mean, are homoscedastic, and are uncorrelated. The design matrix X may contain a column of 1's, in which case a constant is fitted. We denote by x_j^T the j th row of X . When X is of full rank, the *least squares method* leads to the fitted values $\hat{y} = X(X^T X)^{-1} X^T y = Hy$, where H is the so-called *hat matrix*. This matrix describes the orthogonal projection of $\mathbb{R}^n$ onto $C(X)$, the linear

space spanned by the columns of X .

Hoaglin and Welsch (1978) discussed the importance of H in determining leverage points. The diagonal elements of H , denoted by H_{jj} , measure the rate of change in $\hat{y}_j$ as y_j changes. As $0 \leq H_{jj} \leq 1$, the rate of increase in $\hat{y}_j$ is constrained to be no greater than the rate of increase in y_j . Leverage has been generalized to nonlinear regression models. Ross (1987) introduced the *tangent plane leverage* measure and Emerson et al. (1984) defined the *Jacobian leverage* measure. St. Laurent and Cook (1992) showed that the latter measure may exceed 1 (i.e. the rate of change in the predicted response $\hat{y}_j$ may be greater than the rate of change in y_j). A case for which this occurs is said to exhibit *superleverage*.

This paper focus on the problem of *masking* in linear regression. As a rule to determine leverage points, Hoaglin and Welsch (1978) proposed “*design point x_j is a leverage point, if $H_{jj} > 2p/n$.*”

Example 1 Hoaglin and Welsch (1978) quote an example from Draper and Stoneman (1966) with $n = 10$ and $p = 3$. We plot the explanatory variables consisting of specific gravity and the moisture content in Figure 1. The hat matrix for these data is

$$H = \begin{pmatrix} .418 & -.002 & .079 & -.274 & -.046 & .181 & .128 & .222 & .050 & .242 \\ & .242 & .292 & .136 & .243 & .128 & -.041 & .033 & -.035 & .004 \\ & & .417 & -.019 & .273 & .187 & -.126 & .044 & -.153 & .004 \\ & & & .604 & .197 & -.038 & .168 & -.022 & .275 & -.028 \\ & & & & .252 & .111 & -.030 & .019 & -.010 & -.010 \\ & & & & & .148 & .042 & .117 & .012 & .111 \\ & & & & & & .262 & .145 & .277 & .174 \\ & & & & & & & .154 & .120 & .168 \\ & & & & & & & & .315 & .148 \\ & & & & & & & & & .187 \end{pmatrix}$$

The element $H_{44} = 0.604$ exceeds $6/10 = 2p/n$.

The leverage indicator based on the diagonal of H is simple, but can be misleading due to masking. If leverage points occur in groups, the corresponding diagonal values become smaller and may not be spotted by a rule based solely on the size of the diagonal. This poses a challenge in the case where the dimension p becomes large, and no simple picture can reveal the leverage group. Numerical indicators for leverage groups based on H are discussed in Belsley, Kuh and Welsch (1980, p.31) and Cook and Weisberg (1982, Sec.3.6). Let $I = \{i_1, \dots, i_m\} \subset \{1, \dots, n\}$ denote a set of indices. An analysis of the effect of multiple row deletions leads Cook and Weisberg to propose the “*maximum potential $\lambda_{\max}(I)/(1 - \lambda_{\max}(I))$ as the most desirable single number summary.*” In this expression, $\lambda_{\max}(I)$ denotes the largest eigenvalue of the minor H_I of the hat matrix given by the intersection of the rows and columns from I .

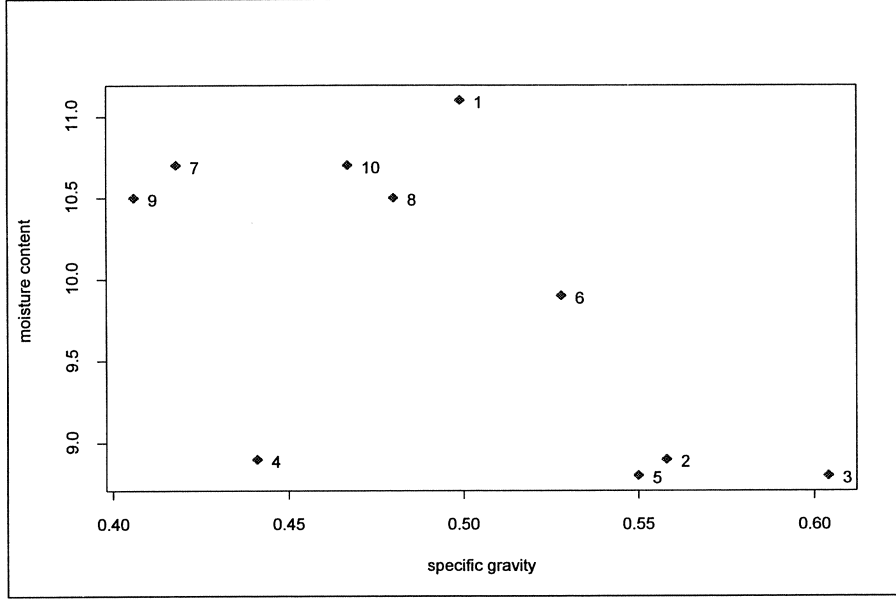


Figure 1: A plot of the variables specific gravity versus moisture content for ten woodbeams. The fourth observation is a leverage point.

The indicator $\lambda_{\max}(I)$ has a simple geometrical interpretation which makes its use even more appealing. Let $\mathcal{E}$ and $\mathcal{F}$ denote two linear subspaces of n -dimensional euclidean space and denote by φ the minimal angle between any two points $e \in \mathcal{E}$ and $f \in \mathcal{F}$. We leave it to the reader to show that this minimum is necessarily achieved for two points e and f that mutually project onto each others direction. If we denote by $P_{\mathcal{E}}$ the orthogonal projection onto $\mathcal{E}$, then $P_{\mathcal{E}}f \propto e$ and vice versa. This shows that the cosine of the minimal angle satisfies

$$(\cos(\varphi))^2 = \text{largest eigenvalue of } P_{\mathcal{E}}P_{\mathcal{F}}P_{\mathcal{E}}.$$

This equation reveals a possible algorithm for solving a minimal angle problem. One starts with an arbitrary $e \in \mathcal{E}$ and alternates projections onto $\mathcal{F}$ and $\mathcal{E}$ followed by normalisations. Applying this argument to the regression problem shows that $\lambda_{\max}(I)$ describes the minimal angle φ between $\mathcal{F} = C(X)$ and $\mathcal{E} = \mathcal{E}_I = \langle e_{i_1}, \dots, e_{i_m} \rangle$, the subspace spanned by the axes with indices in I . In fact, we have

$$\cos(\varphi)^2 = (e^T H e)^2 / ((e^T e)(e^T H^T H e)) = (e^T H e) / (e^T e),$$

where φ is the angle between $e \in \mathcal{E}$ and $H e \in C(X)$. The maximal value of $\cos(\varphi)^2$ is equal to $\lambda_{\max}(I)$.

The least-squares estimator is *equivariant* in the sense that the fitted values $\widehat{y + f} = \hat{y} + f$ for all $f \in C(X)$. In the leverage analysis of the subset indexed by I we check to what degree a perturbation $y \rightarrow y + e$ for $e \in \mathcal{E}_I$ perturbs the fitted values. If the linear subspace $\mathcal{E}_I$ is “nearly parallel” to $C(X)$, that is, if the minimal angle is small, then this perturbation will be large due to equivariance. The minimal angle is, therefore, a leverage diagnostic.

Using an indicator such as $\lambda_{\max}(I)$ poses two kinds of problems. On one hand, the computational challenge is formidable. There are $\binom{n}{m}$ index sets of size m to consider and m should be varied in a reasonable range as well. On the other hand, we need to determine critical values for the indicator. Groups of design points having a value of the leverage indicator larger than the critical value can then be flagged for further consideration. The first problem is the more serious one and will be addressed in Section 2 along with a few comments and a possible solution for the other problem. Section 4 gives some examples.

2 Bounds on eigenvalues

At first sight the proposed leverage indicator seems unpractical, since for each m we must compute $\binom{n}{m}$ eigenvalues. (If $n = 20$ and $m = 5$, for example, there are around 15500 eigenvalues to consider.) However, we are of course merely interested in the largest eigenvalues and these can be picked out quite effectively by using approximations. For symmetric matrices one has the following result.

Theorem 1 (see, for example, Bulirsch and Stoer, 1980, p.385). *The union of the intervals*

$$K_i = \{z \in \mathbb{R} : |z - A_{ii}| \leq \sum_{\substack{k=1 \\ k \neq i}}^m |A_{ik}|\}$$

contains all the eigenvalues of the symmetric $m \times m$ matrix A .

The proof of this result is simple and it tells us that the maximum

$$u(I) = \max_{i \in I} \left(H_{ii} + \sum_{\substack{k \in I \\ k \neq i}} |H_{ik}| \right) \quad (1)$$

is an upper bound on the largest eigenvalue $\lambda_{\max}(I)$ of the minor H_I . Lower bounds on the maximal eigenvalues of minors of H are also easy to find.

Because of their geometrical interpretation, we have for all $e \in \mathcal{E}_I$

$$(e^T H e)/(e^T e) \leq \lambda_{\max}(I). \quad (2)$$

Example 1 (continued). *Inspection of the hat matrix for the wood beam data shows that $|H_{23}|$ followed by $|H_{79}|$, $|H_{49}|$, $|H_{14}|$, and $|H_{35}|$ are the largest off-diagonal elements. The index sets $\{1, 4\}$, $\{4, 9\}$, and $\{1, 4, 9\}$ have large upper bounds, but since the 4th observation has already been flagged, these subsets add little to our understanding. This leaves the pairs $\{2, 3\}$ with upper bound $0.417 + 0.292 = 0.709$ and the triplet $\{2, 3, 5\}$ with upper bound $0.709 + 0.273 = 0.982$ as potentially interesting.*

When $m = 2$, the upper bound on the largest eigenvalue of the minor corresponding to $I = \{i, j\}$ is equal to $\max\{H_{ii}, H_{jj}\} + |H_{ij}|$, whereas the true value is $(H_{ii} + H_{jj})/2 + (H_{ij}^2 + (H_{ii} - H_{jj})^2/4)^{1/2}$. The upper bound is exact if the minor is a diagonal matrix or if the diagonal elements are equal.

Tighter bounds on maximal eigenvalues can be constructed recursively. Let I be a set of indices of size m and let k be an additional index. Any $e \in \mathcal{E}_{I \cup \{k\}}$ can be written as a linear combination $\eta_1 \tilde{e} + \eta_2 e_k$ for some $\tilde{e} \in \mathcal{E}_I$. The angle between e and $C(X)$ has squared cosine value

$$\begin{pmatrix} \eta_1 & \eta_2 \end{pmatrix} \begin{pmatrix} \tilde{e}^T H \tilde{e} & \tilde{e}^T H e_k \\ \tilde{e}^T H e_k & e_k^T H e_k \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix},$$

where we assumed without loss of generality that $\eta_1^2 + \eta_2^2 = 1$. To obtain a lower bound on the squared cosine of the smallest such angle, a good candidate for the choice of $\tilde{e}$ is the eigenvector $e(I) \in \mathcal{E}_I$ which points in the direction of smallest angle with $C(X)$. For linear combinations $e = \eta_1 e(I) + \eta_2 e_k \in \mathcal{E}_{I \cup \{k\}}$ with $\eta_1^2 + \eta_2^2 = 1$, the bound (2) on $\lambda_{\max}(I \cup \{k\})$ is

$$\begin{pmatrix} \eta_1 & \eta_2 \end{pmatrix} \begin{pmatrix} \lambda_{\max}(I) & e(I)^T H e_k \\ e(I)^T H e_k & H_{kk} \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix}.$$

The maximal value of this expression with regard to the choice of η_1 and η_2 , is equal to

$$l(I \cup \{k\}) = \frac{(H_{kk} + \lambda_{\max}(I))/2}{+\sqrt{((e(I)^T H e_k)^2 + (H_{kk} - \lambda_{\max}(I))^2/4)}} \quad (3)$$

and defines a lower bound for $\lambda_{\max}(I \cup \{k\})$.

Since we propose to let the largest upper bounds indicate those sets I with large eigenvalues, we should be sure that there is a correspondance

between large upper bounds and large eigenvalues. The following simple proposition shows how such a relation can be achieved.

Proposition 1 *Denote by $u(I)$, $l(I)$ an upper bound and a lower bound, respectively, for the maximal eigenvalue $\lambda_{\max}(I)$ of index sets of size m . Let I_1 be such a set of indices. If we compute the exact eigenvalues of all index sets I with $u(I) \geq l(I_1)$, we are sure to find among them the largest eigenvalue.*

The proof follows from

$$\lambda_{\max}(I) \leq u(I) < l(I_1) \leq \lambda_{\max}(I_1)$$

for any set I with $u(I) < l(I_1)$. Applying this proposition to the case $m = 2$ leads to the following corollary.

Corollary 1 *If $I_0 = \{i, j\}$ is such that H_{ii}, H_{jj} are the largest diagonal values of H and $|H_{ij}|$ is the largest off-diagonal element in absolute value, then the largest upper bound corresponds to the largest eigenvalue, both being achieved at H_{I_0} .*

Choosing the simple lower bound $\tilde{l}(\{i, j\}) = (H_{ii} + H_{jj})/2 + |H_{ij}|$ we have $u(I) < \tilde{l}(I_0)$ for all $I \neq I_0$, which proves this result. Since eigenvalues of 2×2 minors of H close to unity are possible only if both the diagonal elements and the off-diagonal element of the minor are relatively large, this shows that the upper bound is a useful proxy when $m = 2$. For larger values of m a similar analysis of the quality of the upper bounds is difficult.

3 An algorithm

With regard to the upper bound (1) we propose the following procedure. Let m be a fixed size for the leverage group. Select for each row i of the matrix H the $m + 2$ columns j with the largest absolute values $|H_{ij}|$. Next, calculate the $\binom{m+2}{m}$ possible sums $\sum_j |H_{ij}|$ that have exactly m terms. In this manner, $n \cdot \binom{m+2}{m}$ such sums have been computed when we are done with all the rows. At this stage, we have identified candidates for inclusion in a more detailed analysis. For the $2n$ minors corresponding to the largest absolute row sum, we then calculate the actual maximal eigenvalues numerically.

This puts us in a position to compute the lower bound (3) for all such sets of indices. Let I be a set of size m for which $\lambda_{\max}(I)$ and $e(I)$ are known. First, we search for good candidate indices k to be added. This we do by using the above procedure for the upper bounds with m increased to

$m = 1$	0.604	4		
	0.418	1		
	0.417	3		
$m = 2$	0.800	1,4	9	
	0.770	4,9	1	
	0.693	4,5	9	
$m = 3$	0.905	4,7,9	1	0.903
	0.876	1,4,9	5	
	0.859	1,4,5	9	
$m = 4$	0.949	3,4,7,9		0.9432
	0.940	1,4,7,9		
	0.935	2,3,5,6		

Table 1: Application of the above algorithm to the data on wood beams. The table shows the largest maximal eigenvalues for subsets of size $m = 1, 2, 3, 4$. For $m = 2, 3$ the fourth column shows the additional index used in computing the lower bound for the next stage. The largest of these lower bounds is given in the fifth column for $m = 3, 4$.

$m + 1$. This way, we select a certain number of sets of size $m + 1$. If none of them is of the form $I \cup \{k\}$, we do nothing. If one or several are of that form, we choose that index with the largest upper bound and compute (3). The computational complexity of this algorithm is not overwhelming. We implemented it on a 386-based PC with satisfactory results.

Example 1 (continued). *Approaching the data on wood beams in this manner leads to the results shown in Table 1.*

In all instances, the lower bound is excellent and we are sure to have identified the largest maximal eigenvalues of $m \times m$ minors for all the values of m considered.

The example shows the typical behavior of $\lambda_{\max}(I)$. As the size of I increases, so does the largest eigenvalue. Furthermore, subsets that have a large value of the diagnostic lead to many supersets that also have large values. We need to flag a subset as being leverage as soon as all its members are assembled. Subsequently, we should no longer consider superset of the flagged group. To test whether a leverage group is complete, we must compute the increase in the largest eigenvalue when adding new points to the group. If a few of these additions lead to a disproportionally large increase, not all members have been included. If on the other hand all or most of the additions lead to a largest eigenvalue indicating leverage, then the group is already complete and the supersets are of no interest.

It is easiest to implement these ideas if we have available a cutoff value for determining whether or not an eigenvalue is large enough. Based on our experience with examples, we propose

$$\gamma_m = \frac{n^2 - 4p}{n(n-2)} - \frac{n-2p}{m(n-2)},$$

where m is the dimension of the minor, n is the number of observations and p is the dimension of the regression parameter (including the constant). Note that this function of m interpolates between $\gamma_1 = 2p/n$, the Hoaglin and Welsch (1978) cutoff, and $\gamma_{n/2} = 1$.

With regard to changes in the value of m these leads to the following recursive procedure. Start with $m = 1$ and apply the following rules.

1. If $\lambda_{i_1, \dots, i_m} < \gamma_m$ for each $(i_1, \dots, i_m)$, then go to (3).
2. If there is a group of design points $(x_{i_1}, \dots, x_{i_m})$ with $\lambda_{i_1, \dots, i_m} > \gamma_m$ we distinguish two cases:
 - (a) there is an additional point which should be included in the group, in which case go to (3);
 - (b) the group forms its own cluster, in which case we flag $(y_{i_1}, \dots, y_{i_m})$ as a high leverage subset and then go to (3). Once a high leverage group is determined, we delete it from further consideration.
3. Increase the value of m and go to (1).

To choose between (2a) and (2b), compute $\lambda_{\max}(i_1, \dots, i_m, k)$ for $k \neq i_1, \dots, i_m$. We select (2a), if exactly one of these eigenvalues exceeds γ_{m+1} . If none or more than one exceed γ_{m+1} , we select (2b). Whenever (2b) is chosen, the eigenvalues $\lambda_{\max}(i_1, \dots, i_m, k)$ are typically larger than or very close to γ_{m+1} . But, since our algorithm considers a single additional design point, it does not work satisfactorily in some situations. For the more limited objective, however, of identifying small clusters of outlying design points, our simple strategy suffices.

Example 1 (continued). *The application of this procedure to the example of Section 1 leads to the following comments. The fourth observation exceeds the cutoff of $2p/n = 0.6$. If we consider all subsets of size two that include the fourth point, we find $\lambda_{\max}(1, 4) = 0.80$, $\lambda_{\max}(2, 4) = 0.65$, $\lambda_m(3, 4) = 0.60$, $\lambda_m(4, 5) = 0.69$, $\lambda_m(4, 6) = 0.61$, $\lambda_m(4, 7) = 0.67$, $\lambda_m(4, 8) = 0.61$, $\lambda_m(4, 9) = 0.77$ and $\lambda_m(4, 10) = 0.61$. No value exceeds γ_2 , which leads us to flag the fourth point as high leverage. No other subgroups are found.*

$m = 1$	0.525	7		
	0.471	16		
	0.466	10		
$m = 2$	0.775	7,9	3	
	0.680	11,12	2	
	0.673	10,13	17	
$m = 3$	0.868	7,9,18	3	0.8675
	0.822	7,9,17	3	
	0.813	7,9,13	18	
$m = 4$	0.960	4,6,8,19	11	0.9589
	0.903	7,9,13,18	5	
	0.902	7,9,15,18	3	
$m = 5$	0.979	4,6,8,11,19		0.9790
	0.964	4,6,8,16,19		
	0.942	5,7,9,13,18		

Table 2: Application of the above algorithm to Rousseeuw's data. The table shows the largest maximal eigenvalues for subsets of size $m = 1, 2, 3, 4, 5$. For $m = 2, 3, 4$ the fourth column shows the additional index used in computing the lower bound for the next stage. The largest of these lower bounds is given in the fifth column for $m = 3, 4, 5$. Note that the outliers make a sudden appearance when $m = 4$.

4 Two examples

The first example is an artificial one constructed by Rousseeuw (1984) to illustrate the advantages of the least median squares estimator. Rousseeuw imputed in a clever manner four outliers into the data set, namely the observations 4, 6, 8 and 19. In a dynamic 3-dimensional graphic display, these points form a clearly visible high leverage subset, but any simplistic analysis of the hat matrix fails. We shall now apply our algorithm. The sequence of cutoff values is $\gamma_1 = 0.600$, $\gamma_2 = 0.822$, $\gamma_3 = 0.896$, $\gamma_4 = 0.9333$, $\gamma_5 = 0.956$, $\gamma_6 = 0.970$, $\gamma_7 = 0.981$, $\gamma_8 = 0.989$, $\gamma_9 = 0.995$, $\gamma_{10} = 1$. The largest diagonal value is $H_{77} = 0.525 < \gamma_1$. For larger values of m , the large eigenvalues are $\lambda_{\max}(7, 9) = 0.775$, $\lambda_{\max}(7, 9, 18) = 0.868$ and $\lambda_{\max}(4, 6, 8, 19) = 0.960$. Table 2 summarizes the results.

The algorithm flags the four imputed outliers as forming a leverage group, since $\lambda_{\max}(4, 6, 8, 19, k)$ is large for all $k \neq 4, 6, 8, 19$. For $m = 5$, Table 2 shows that the largest eigenvalue, without considering the indices 4, 6, 8 and 19, is $\lambda_{\max}(5, 7, 9, 13, 18) = 0.942$, which is smaller than γ_5 . For $m > 5$ no further candidates are found either.

X_1	X_2	X_3	X_4	X_5	Y
1125	232	7160	85.9	8905	1.5563
920	268	8804	86.5	7388	0.8976
835	271	8108	85.2	5348	0.7482
1000	237	6370	83.8	8056	0.7160
1150	192	6441	82.1	6960	0.3160
990	202	5154	79.2	5690	0.3617
840	184	5896	81.2	6932	0.1139
650	200	5336	80.6	5400	0.1139
640	180	5041	78.4	3177	−0.2218
583	165	5012	79.3	4461	−0.1549
570	151	4825	78.7	3901	0.0000
570	171	4391	78.0	5002	0.0000
510	243	4320	72.3	4665	−0.0969
555	147	3709	74.9	4642	−0.2218
460	286	3969	74.4	4840	−0.3979
275	198	3558	72.5	4479	−0.1549
510	196	4361	57.7	4200	−0.2218
165	210	3301	71.8	3410	−0.3979
244	327	2964	72.5	3360	−0.5229
79	334	2777	71.9	2599	−0.0458

Table 3: Moore’s data.

As a second example, we use the result of a laboratory experiment performed by Moore (1975). The data are reproduced in Table 3. The independent variable Y is the oxygen demand in dairy wastes. The explanatory variables are X_1 = biological oxygen demand (mg/liter), X_2 = total Kjeldahl nitrogen (mg/liter), X_3 = total solids (mg/liter), X_4 = total volatile solids, a component of X_3 (mg/liter) and X_5 = chemical oxygen demand (mg/liter). Table 4 shows the results of applying our algorithm. The design point 17 and the points (2,3) could be high leverage.

The sequence of calibration points is the same as for Rousseeuw’s data. The largest diagonal element of the hat matrix is $H_{17,17} = 0.918 > \gamma_1$. The computation of $\lambda_{j,17}$ where $j \neq 17$ produces values which are much bigger than γ_2 . Thus the observation y_{17} is distinctly high leverage, forming a group by itself. For $m = 2$, the largest eigenvalue is $\lambda_{\max}(2, 3) = 0.877$. It exceeds γ_2 and since all the values $\lambda_{\max}(2, 3, k)$ ($k \neq 2, 3$) are large, the subset $\{2, 3\}$ is also flagged. These two observations are jointly high leverage. The indices 2, 3 and 17 are left out in the sequel. For $m = 3$, the largest eigenvalue is $\lambda_{\max}(5, 19, 20) = 0.887 < \gamma_3$. For $m = 4$ we

$m = 1$	0.918	17		
	0.502	2		
	0.485	3		
$m = 2$	0.877	2,3	9	
	0.752	19,20	15	
	0.637	3,9	2	
$m = 3$	0.887	15,19,20	18	0.7554
	0.788	13,19,20	15	
	0.778	18,19,20	15	
$m = 4$	0.927	13,15,19,20		0.9280
	0.912	15,18,19,20		
	0.904	6,15,19,20		

Table 4: Application of the above algorithm to Moore's data. The table shows the largest maximal eigenvalues for subsets of size $m = 1, 2, 3, 4$. For $m = 2, 3$ the fourth column shows the additional index used in computing the lower bound for the next stage. The largest of these lower bounds is given in the fifth column for $m = 3, 4$. The exhibit for $m > 1$ excludes the index $i = 17$ and for $m > 2$ we also delete the indices $i = 2, 3$.

have $\lambda_{\max}(13, 15, 19, 20) = 0.927 < \gamma_4 = 0.933$. For $m = 5$ we find $\lambda_{\max}(13, 15, 18, 19, 20) = 0.9540 < \gamma_5 = 0.9560$ and, finally, for $m = 6$ we have $\lambda_{\max}(7, 13, 15, 18, 19, 20) = 0.960 < \gamma_6 = 0.970$. No further leverage subsets are detected.

This same data set has been analysed by Hadi and Wells (1990) with regard to the effect of multiple row deletions on the collinearity. Their measure identifies 17 and $\{13, 17\}$ as being influential with regard to collinearity. Had we not deleted the 17th point from further considerations, the same pair $\{13, 17\}$ would occupy the top spot in the exhibit for $m = 2$.

5 Concluding remarks

We have discussed the use of $\lambda_{\max}(I)$, the maximum eigenvalue of the minor of the hat matrix H corresponding to the set of indices I , as a diagnostic for leverage. We gave a geometrical interpretation of this number and discussed a heuristic algorithm that makes the computation feasible. We also proposed a solution to the problem of setting a critical value for judging the size of $\lambda_{\max}(I)$, and used this choice to build an algorithm for identifying leverage subsets. The lower bound that can be computed recursively is of extremely high quality, but it does not save us from difficulties, because our simple upper bound is crude. The proposition of Section 2 asks us

to test all subsets with an upper bound that is bigger than the maximal value of the lower bound. This works well for values of m up to about six. However, experience shows that while there is certainly no one-to-one correspondance between large upper bounds and large eigenvalues, there is a surprisingly tight relationship even for large values of m .

Résumé: Hoaglin et Welsch (1978) discutent l'emploi de la matrice H en régression et en ANOVA pour trouver les points leviers ("leverage points"). Un point levier est n'importe quelle observation dont l'élément diagonal de H est grand. Dans cet article, nous essayons d'utiliser également les éléments de H se trouvant hors de la diagonale. Notre approche nécessite le calcul de certaines valeurs propres de sous-matrices de H . Elle permet l'identification d'ensembles d'observations dits sous-ensembles leviers ("leverage subsets") qui auraient été oubliés si seuls les éléments diagonaux avaient été considérés.

Mots clés: Matrice H , points leviers, sous-ensembles leviers.

References

- [1] Belsley, D.A., Kuh, E. and Welsch, R.E. (1980), *Regression Diagnostics*, John Wiley, New York.
- [2] Bulirsch, R. and Stoer, J. (1980), *Introduction to Numerical Analysis*, Springer-Verlag, Heidelberg.
- [3] Cook, R.D. and Weisberg, S. (1982), *Residuals and Influence in Regression*, Chapman and Hall, London.
- [4] Emerson, J.D., Hoaglin, D.C. and Kempthorne, P.J. (1984), "Leverage in Least Squares Additive-Plus-Multiplicative Fits for Two-Way Tables", *J. Amer. Statist. Assoc.*, 79, 329–335.
- [5] Hadi, A.S. and Wells, M.T. (1990), "Assessing the Effects of Multiple Rows on the Condition Number of a Matrix," *J. Amer. Statist. Assoc.*, 85, 786–792.
- [6] Hoaglin, D.C. and Welsch, R.E. (1978), "The Hat Matrix in Regression and ANOVA", *The American Statistician*, 32, 17-22.
- [7] Moore, J.A. (1975), "Total Biochemical Oxygen Demand of Animal Manure", Unpublished PhD Dissertation, University of Minnesota.
- [8] St. Laurent, R.T. and Cook, R.D. (1992), "Leverage and Superleverage in Nonlinear Models", *J. Amer. Statist. Assoc.*, 87, 985–990.
- [9] Ross, W.H. (1987), "The Geometry of Case Deletion and the Assessment of Influence in Nonlinear Regression", *Can. J. Statist.*, 15, 91–103.
- [10] Rousseeuw, P.J. and Leroy, A.M. (1987), *Robust Regression and Outlier Detection*, John Wiley, New York.

Markov Chains and Vegetation Monitoring

Helene H. Wagner and Otto Wildi

*Swiss Federal Institute for Forest, Snow and Landscape Research,
Birmensdorf, Switzerland*

Abstract: What can the application of Markov chains, in comparison to more traditional multivariate techniques, contribute to trend analysis in the context of short multi-species time-series that comprise a large amount of fluctuation? Two selected methods were compared by simulation and by analysis of real-life data from a dry meadow monitoring program of the Canton Aargau (Switzerland). The Markovian method accurately distinguished simulated trend and noise, whereas the regression-based approach effectively measured the amount of change only. Combination of the two methods on plot and site level showed general trend in five sites, predominant noise in one, divergent development of different parts of the meadows in two and ambiguous results for one site. To avoid misinterpretation, change should always be examined in terms of the two qualities, trend and rate. Markovian models are a powerful tool to assess trend in vegetation change. To make their whole potential available for prediction, alternative sampling strategies or transformations of the data may have to be evaluated.

Key words: Dry meadow, multivariate data, noise, time series, trend.

1 Introduction

In the context of environmental change, there is a rising number of vegetation monitoring programs going on. Data on plant species composition are sampled in an environment where neither the occurrence of trend nor its direction or rate are known in advance. Analysis should not only detect systematic change, but accurately describe and predict it in terms of individual species development.

The Canton Aargau (Switzerland) has been running a monitoring program for dry meadows since 1988. Its aim is to watch the development, in terms of changing plant species composition, of the best preserved dry meadows of the canton. The scheme includes yearly observations of species abundance, i.e. frequency of occurrence, in 47 permanent plots at nine selected locations. In 1995, we were asked to perform trend analysis and to assess the program's efficiency. In order to answer this question, the data had to be scrutinised for the existence, the rate and the direction of any observed trend.

The data are multivariate time series, where relevés (observation vectors) consist of a specific combination of abundances of a high number of species (variables). Their comparative development, i.e. gains and losses from one population to the other, conveys the central ecological information. Series are short (7-8 steps) and contain much fluctuation. The number of variables is excessively high and considerably larger than the number of relevés.

Multivariate analysis methods are well developed in quantitative plant ecology (see e.g. Jongman et al., 1995). Ordination techniques (variants of Principal Component Analysis (PCA) or Multi-Dimensional Scaling (MDS)) are used for data reduction and graphical representation. They enable us to explore graphically the development over time, but trend confirmation and prediction are not possible. Wildi (1988) presented a regression-based method to model linear trend in a short vegetation time series. It measures trend by comparison of states, but extrapolation is not possible.

The possibilities of modelling plant-plant substitution as a stochastic process, especially with Markov chains, have been discussed by Fekete (1985) and Usher (1981). Competition is thought to govern the substitution of an individual by an individual of another species. The substitution of population *a* by population *b* within a certain time interval can be described as a probability. All succession probabilities within a certain phase are summarised in a transition matrix, by which the process is defined entirely. It can then be described by Markov chains.

The approach seems promising as it, in principle, allows the description and prediction of change. There are important limitations, though. Firstly, the appropriateness of the necessary assumptions is often questionable. Secondly, as Usher (1981) pointed out, the definition of states is fundamental to the later development and interpretation of a model. Transition probabilities are typically derived from observing the development of a large number of plots, defining a state either in terms of the dominant species or as one type of relevés. Both possibilities are problematic, as the specific combination of species abundances has to be strongly generalised

in order to have enough information to estimate all transition probabilities. A possible solution to the problems of estimating transitions probabilities even for a large number of states based on relatively few, survey type data, and of Markovity testing was presented by Orlóci et al. (1993).

Our aim was to investigate what the application of Markov chains, in comparison to more traditional multivariate techniques, can contribute to trend analysis in the context of short vegetation time series with a large amount of fluctuation. The two methods by Wildi (1988) and by Orlóci et al. (1993) were compared by simulation and by analysis of real-life data from the dry meadow monitoring program of the Canton Aargau.

2 The dry meadow data

At the beginning of the monitoring program experts had selected a group of 9 dry meadows, based on an inventory established in 1982. The selection was supposed to represent the “best” dry meadows of the Canton Aargau. At each site, large samples of plots were taken. With a procedure described by Wildi (1990), a group of 5-6 plots per site was selected for permanent observation so that they would represent a maximum of variation within the site. For each permanent plot, species abundance was recorded annually from 1989 (some from 1988) to 1995, counting for each species in how many of 6 segments of a hexagon of 2.14 m² area it occurred. The notation is as follows:

p	permanent plot ($p = 1, 2, \dots, 47$)
$\mathbf{X}$	matrix of abundance scores
i	species ($i = 1, 2, \dots, I_p$)
t	time point ($t = 1, 2, \dots, T_p$)
x	abundance score ($x = 0, 1, \dots, 6$)

The data were organised per plot in a matrix $\mathbf{X}_p$, where the element x_{pit} is the abundance score x of species i at time point t :

$$\mathbf{X}_p = \begin{bmatrix} x_{p11} & \dots & x_{p1t} & \dots & x_{p1T_p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{pi1} & \dots & x_{pit} & \dots & x_{piT_p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{pI_p1} & \dots & x_{pI_pt} & \dots & x_{pI_pT_p} \end{bmatrix}$$

The series consisted of 7 to 8 relevés (T_p). Total species numbers per plot over the whole period (I_p) varied between 33 and 79. The numbers of species per plot and year were considerably lower (11-60), indicating a large amount of fluctuation.

At one site, data from 1994 were missing as the meadow had been mown before observation. As a result of data quality analysis, recordings from the first year of observation were excluded from trend analysis for three sites.

3 Simulated series

To compare the two methods, several types of simulated series with known properties were analysed and results contrasted. Relevés were simulated along a one-dimensional time gradient as proposed by Lagonegro (1984, see also Wildi, 1986, for a thorough account incl. references). It was assumed that species abundance increases and decreases in time, which was modelled using a normal distribution function.

To approximate the observed relevés, 40 species were introduced and their scores were limited to the range of the original data, i.e. from 0 to 6. The location μ was randomised and a large variance σ^2 assumed. Six combinations of different gradient lengths (long trend: 3.3σ , short trend: 1.7σ) and noise levels (0%, 50% and 90%) were simulated. A systematic sample of 7 relevés was taken from each gradient. Two random samples were taken from the long gradient with no noise.

4 The method proposed by Wildi (1988)

Wildi (1988) introduced a method to measure linear trend within short, multivariate vegetation time series in a situation where changes in species composition are mainly continuous and not drastic.

Assumptions concern the linearity of trend in species scores and the absence or irrelevance of any cyclic component. Although cyclic phenomena are common among climax vegetation types like some forests or heaths, where stands collapse after reaching a certain age, they were supposed to play a negligible role in the case of the dry meadows concerned. Likewise, species reaction to a gradient tends to be bell-shaped rather than linear (see Gauch, 1982), but linearity of trend in species scores was assumed here because the time span was short and it was known that no drastic changes had occurred.

Analysis was performed using the software MULVA-5. For each plot p , regression parameters $b_{0_{pi}}$ and $b_{1_{pi}}$ were estimated for every species i by linear least squares regression of abundance scores x_{pi} on time t . From this, expected scores y_{pit} were calculated for $t = 1, 2, \dots, T_p$ according to

$$y_{pit} = b_{0_{pi}} + b_{1_{pi}} \cdot t$$

As in the original paper, predicted values y_{pit} were bounded to the range

of the original abundance values, i.e. from 0 to 6. To give more weight to rarer species, the fitted values y_{pit} were root-transformed according to

$$\tilde{y}_{pit} = \sqrt{y_{pit}}$$

To make relevés with different sums $\sum_i \tilde{y}_{pit}$ comparable, relevé vectors of fitted values were normalised according to

$$\sqrt{\sum_i \tilde{y}_{pit}^2} = 1$$

The similarity of the first and last fitted relevés (at $t = 1$ and $t = T_p$) was determined by calculating the similarity ratio SR (see Ball, 1966):

$$SR_{1T} = \frac{\sum_i \tilde{y}_{i1} \cdot \tilde{y}_{iT}}{\sum_i \tilde{y}_{i1}^2 + \sum_i \tilde{y}_{iT}^2 - \sum_i \tilde{y}_{i1} \cdot \tilde{y}_{iT}}$$

For simplicity, the suffix p was omitted from the formula. This is an extension to quantitative data of Jaccard's index (see Jaccard, 1912):

$$SJ_{jk} = \frac{c}{A + B - c}$$

where A and B are the numbers of species in relevés j and k , and c is the number of species shared. Unlike a correlation coefficient, SR is not influenced by the number of species with zero scores in both relevés (for a discussion of similarity measures in the context of vegetation data see Gauch, 1982). SR has a range from 0 to 1, and levels of variation are known from the literature. Wildi (1988) argues that in comparing plant communities, a drop of SR below about 0.5 generally indicates a fundamental change of the vegetation type. A similarity smaller than 0.4 was supposed to indicate substantial change, whereas a value between 0.4 and 0.7 was rated as moderate and one larger than 0.7 as little change.

5 The method proposed by Orlóci et al. (1993)

Orlóci et al. (1993) presented a technique to determine transition probability matrices from survey-type vegetation data and to perform distribution-free significance tests for the null hypotheses ' H_0 : series is undirected' and ' H_0 : series is Markov with m -step transition matrix'. Instead of generalised overall states of a plot, the states of a small portion of ground surface are considered and defined in terms of a single species occupying it. The observed partition of ground among all competing species at a certain time is used as an initial probability vector.

Formal conditions of the model are the Markov property and stationarity, i.e. transition probabilities should be independent of process history and of time, and convergence of the series is implied. These are not general principles in vegetation change (see e.g. Fekete, 1985), but they might be approximated in the present case, as no impact is known to have happened during the observed time span and cyclic changes due to the ageing of stands are not expected. Stationarity and Markovity at step size 1 are tested empirically.

When defining the states, we have to assume that all possible competing species are incorporated into the model and that the data completely and accurately describe the use of the resource that governs their competition. While we suppose that species compete for ground occupation, the system is of course not closed and resource shares are not necessarily reflected by observed frequencies. When constructing the transition matrix, the distribution of gain or loss among competing species is simplified to be proportional to population size.

Ground cover $\tilde{x}_{pit}$ of plot p by species i at time t was modelled by root-transformation of the original score x_{pit} :

$$\tilde{x}_{pit} = \sqrt{x_{pit}}$$

assuming that rarer species (often forbs, shrubs or trees) occupy a larger area per individual than frequent ones (often grasses). The relevé vectors $\tilde{X}_{pt}$ of the transformed values $\tilde{x}_{pit}$ at time t were transformed into stochastic vectors A_{pt} with the elements a_{pit} representing the proportion of ground surface occupied by species i at time t .

$$a_{pit} = \frac{\tilde{x}_{pit}}{\sum_i \tilde{x}_{pit}}$$

The following steps of estimating transition probabilities and hypothesis testing were done as proposed by Orlóci et al. (1993), applying their software “FitMarko”. For further details and discussion, cf. to the original paper. A separate transition matrix was estimated for every plot and for plot-wise averaged series. For simplicity, the suffix p for plot is omitted in the following.

There are three possible developments of species i from time point t to the next: i can either gain ground ($a_{i(t+1)} - a_{it} > 0$), loose ground ($a_{i(t+1)} - a_{it} < 0$) or remain unchanged ($a_{i(t+1)} - a_{it} = 0$).

Gain and loss from species i to h from time t to $(t+1)$ was modelled as:

$$\begin{aligned} DEV(i) &= \left| a_{i(t+1)} - a_{it} \right| \cdot \frac{a_{h(t+1)}}{a_{\cdot(t+1)}} && \text{for } h \neq i \text{ (same for loss and gain)} \\ DEV(i) &= a_{i(t+1)} && \text{for } h = i \end{aligned}$$

assuming that gain from or loss to i by h is proportional to the ground occupied by h at time $(t + 1)$.

The one step transition matrix $\mathbf{P}_t^1$ was estimated based on vector pairs A_t, A_{t+1} as follows: $DEV(i)$ -values were attributed to line i in case of loss or to column i in case of gain. $DEV(i)$ -values were then summed cell-wise to obtain $p_{t_{ih}}^1$'s. This was repeated for each time-interval $[t, t + 1]$ to obtain $\mathbf{P}_1, \mathbf{P}_2, \dots, \mathbf{P}_{T-1}$. These were averaged cell-wise:

$$p_{ih}^1 = \frac{\sum_{t=1}^{T-1} p_{t_{ih}}^1}{T - 1}$$

and then transformed row-wise to fulfil the condition

$$\sum_h p_{ih}^1 = 1$$

leading to the one-step transition matrix $\mathbf{P}^1$. Its elements describe the probability at which ground occupied by species i is overtaken by species h in one time step.

To fit the model, a set of Markov relevés M_t ($t = 1, \dots, T$) was generated by the formula

$$M_{t+1} = M_t \mathbf{P}^1$$

starting with $M_1 = A_1$ as the initial probability vector. Euclidean distances

$$d_{A_{jk}} = \sqrt{\sum_i (a_{ij} - a_{ik})^2}$$

were calculated for all possible vectors pairs A_j, A_k at time points $t = j$ and $t = k$ and similarly $d_{M_{jk}}$ among Markov relevés M_t , leading to the distance matrices $\mathbf{D}_A$ and $\mathbf{D}_M$.

Discordance between $\mathbf{D}_A$ and $\mathbf{D}_M$ was expressed in terms of a stress value σ , which is proportional to the closeness of the model and the observed series. It is defined as:

$$\sigma(\mathbf{D}_A; \mathbf{D}_M) = \sqrt{\frac{\sum_{j < k} (d_{A_{jk}} - d_{M_{jk}})^2}{\sum_{j < k} (d_{A_{jk}} + d_{M_{jk}})^2}}$$

A stationary Markov chain is completely defined by the initial probability vector and the transition matrix $\mathbf{P}$. It is therefore sufficient to show that the series is stationary, i.e. has a constant transition matrix, to show Markovity (see. Orlóci et al., 1993, for discussion). As proposed in the

Simulated data sets		Method by Wildi		Method by Orlóci et al.	
Gradient type	noise level	similarity value SR	amount of change	approx. p -value	directed?
Long	0%	0.32	substantial	0.00	yes
	50%	0.32	substantial	0.00	yes
	90%	0.32	substantial	0.00	yes
Short	0%	0.90	little	0.00	yes
	50%	0.90	little	0.00	yes
	90%	0.91	little	0.00	yes
Random samples	0%	0.74	little	0.12	no
	0%	0.99	little	0.75	no

Table 1: Trend analysis for simulated relevés.

original paper, stationarity was checked with a bootstrap test on the stress value σ . A sampling distribution for σ under the null hypothesis ‘ H_0 : series is undirected’ was simulated. The observed relevés were randomly permuted in order to generate different transition probabilities for every time step, and a simulated stress value σ_r^* was calculated for each replication r ($r = 1, 2, \dots, 100$). The proportion of $\sigma_r^* < \sigma(\mathbf{D}_A; \mathbf{D}_M)$ yields an approximate p -value for H_0 . If H_0 was rejected, the alternative hypothesis ‘ H_1 : series is directed’ was accepted. Transition probabilities were assumed approximately constant and the series therefore describable by a stationary Markov process. H_0 was definitively rejected on a significance level of 0.05 and kept by at a p -value > 0.1 , while for a p -value between 0.05 and 0.1, stationarity was considered questionable.

6 Results from simulation

To compare the two methods, several types of simulated series with known properties were analysed and results contrasted (see Table 1). The multivariate method by Wildi (1988) clearly distinguished between long and short gradient, but not between the random samples and a short gradient. In contrast, the Markovian method by Orlóci et al. (1993) readily discriminated trend from the random samples, but failed to mark a difference between long and short gradient. Neither method was much affected by random noise in species scores, even up to a level of 90% noise.

The two methods obviously measured different qualities of change. While we interpreted the similarity ratio SR from the multivariate method as a measure for the amount of change, the p -value for ‘ H_0 : series is undirected’ was interpreted as a measure for the probability of existence of directed change, i.e. trend.

Type	Site	Number of plots with				Averaged series trended?
		moderate trend	weak trend	evt. trend	noise	
1	Oberhof		5			yes
	Effingen		4	1		yes
	Küttigen Judenhalde		4	1		yes
	Gansingen		3	2		yes
2	Küttigen Egghübel		3		2	yes
3	Hornussen		2	1	3	evt
4	Untersiggenthal	2		2	1	no
	Densbüren		2	2	1	no
5	Sarmenstorf		1	1	4	no

Table 2: Plot results and existence of trend in site-wise averaged series.

7 Results from the dry meadow data

On plot level, similarity scores SR were generally high, indicating low change rates. Only two plots reached values < 0.7 , which implied moderate change, while all other plots showed but little change. Significance values for the null hypothesis ‘ H_0 : series is undirected’ showed a different picture. 26 of 47 plots reached approximate p -values ≤ 0.04 and were classified as trended. Another 10 obtained values between 0.05 and 0.09, indicating possible trend, while 11 plots showed p -values ≥ 0.1 and were classified as lacking trend.

Comparatively low similarity scores, indicating much change, went together with very low probability for trend and vice versa. The two variables were not significantly correlated (Spearman rank correlation $r_s = 0.19$, $p = 0.19$). Of the four observed combinations of classification, “trended, moderate change” was interpreted as a moderate trend (2 plots), “trended, little change” as a weak trend (24 plots), “possibly trended, little change” as a possible weak trend (10 plots) and “not trended, little change” as predominant noise (11 plots).

In addition, we tested site-wise averaged series for trend and contrasted the results with plot classification (see Table 2). In order to interpret the respective findings for plots and averaged series we considered the actual direction of observed trends.

Figure 1 shows extrapolated Markov series for plots from sites Oberhof and Densbüren in an ordination by the second and third axes of Correspondence Analysis (CA). This projection best captured the comparative development of the series, while the first axes mainly comprised the difference between plots.

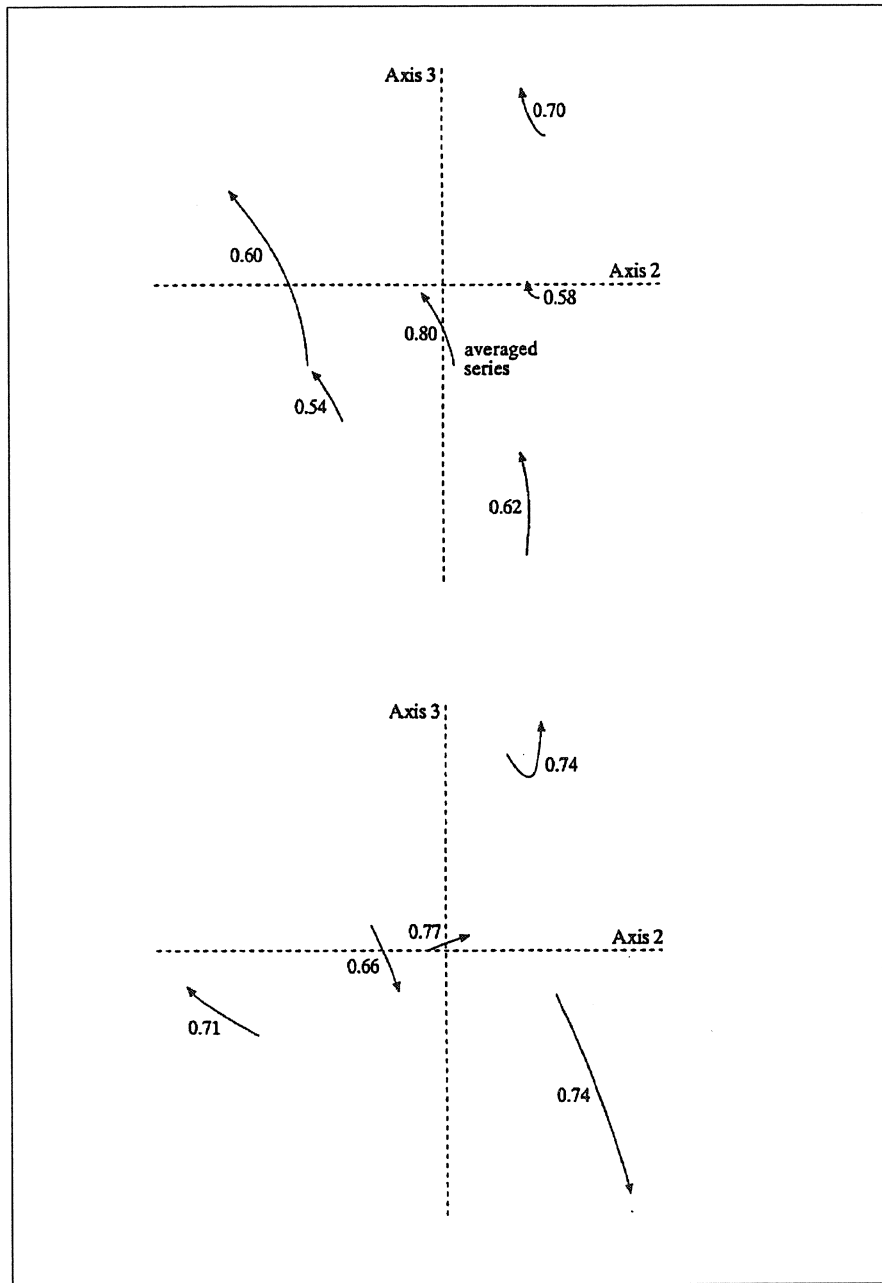


Figure 1: Ordination of extrapolated Markov series for plots and averaged data of the site Oberhof (above) and for plots of the site Densbüren (below). As distances are distorted, similarity values SR of first and last Markov relevés are indicated.

The ordination for Oberhof clearly showed parallel trajectories for plots and the averaged series. In contrast, the plots from Densbüren were clearly diverging (for Densbüren, the averaged series was not included as it had not proved to be directed). According to Austin (1977), the first pattern is typical for a general trend, while the second implies the lack of trend or the existence of site-dependent trends.

The specific combinations of ratings on plot and site level were therefore interpreted as follows:

- Type 1: The plots as well as the averaged series showed a significant trend at least at the 90% level. Plots seemed to develop in a parallel direction. This interpretation was confirmed by the graphical analysis for the site Oberhof (see Figure 1 above). It was supposed to reflect an adaptation process that concerned the whole site. The rate was generally slow.
- Type 2: The situation according to plots was ambiguous, but the averaged series showed a significant trend (at the 95% level). Trend appeared to be covered by much noise and amplified through averaging. A weak trend was presumed.
- Type 3: Neither plots nor the averaged series provided clear results. The further development will have to be investigated to reach a better founded conclusion.
- Type 4: According to the plots, the situation was ambiguous, and the averaged series showed no trend. Different processes seemed to be active in different parts of the site with not much parallel development. This interpretation was confirmed by the graphical analysis for the site Densbüren (see Figure 1 below).
- Type 5: Neither plots nor the averaged series generally showed a significant trend. Natural, random fluctuation from year to year seemed to dominate.

8 Discussion

Using the method by Orlóci et al. (1993), several problems were solved or at least mitigated that have hindered the application of Markov chains to succession modelling. Usher (1981) listed seven areas where the Markovian models are “either difficult, impossible, or unrealistic from a biological point-of-view”. These are: (1) the difficulty of dividing a continuum into discrete states; (2) the large amount of data needed to estimate all possible transition probabilities; (3) the assumptions of Markovity and (4) of stationarity; (5) the lacking of spatial effects in Markovian models; (6) the condition that the data must be fully representative of all the species; and (7) the more philosophical question of whether a model is necessary at all,

once we had enough information to build a good Markovian model which integrated problems of non-stationarity and dependence on more than the last state and we knew that no new species would immigrate?

The method by Orlóci et al. (1993) solves the first problem by considering the states of a small portion of a plot instead of the states of the entire plot. It theoretically allows to estimate a transition matrix based on two relevés, which resolves problem (2). They also developed means to test for Markovity and stationarity empirically. Stationarity and, implicitly, Markovity have been tested for the dry meadow data. Stationarity was assumed for 55% of the investigated series, for another 20% it was likely to apply.

On the other hand, the fifth argument that specific patterns of spread can not be integrated and spatial relationships are generally ignored still holds. The same applies to point (6) that extinction can be predicted, but not immigration. These two problems remain unsolved and restrict application and interpretation of the method.

Point (7) should be discussed in a specific context, as efficiency depends on goals and cost. It becomes less questionable in the present case as a relatively simple model, assuming both stationarity and Markovity, can be used and transition probabilities can be estimated based on few relevés. However, to answer the question of efficiency, we would want to assess the quality of prediction.

The inherent possibility of prediction is of course the main advantage of a Markovian model compared to alternative multivariate methods. It is also the most vulnerable part, as the dependence on assumptions can cause major errors once the Markov chain is extrapolated beyond the observed time span.

In modelling plant-plant substitution as a Markov process we assumed that plants competed for one resource. The resource shares were described in terms of observed frequency counts, which were then transformed to describe a state. We further assumed that gain or loss by one population is distributed among all others proportionally to their size, as proposed by Orlóci et al. (1993).

It seems promising to determine the fixed probability vector A_f (which fulfills the condition $A_f \mathbf{P} = A_f$) of the transition matrix $\mathbf{P}$ to estimate the repartition of resource among species at the equilibrium state. Ecological interpretation could then be achieved by using additional synecological and ecological information about species, as proposed by van der Maarel and Werger (1978). But as the estimation of the transition probabilities depends on prior modelling of competition based on observed abundances, the effect of alternative sampling strategies and transformations should be considered.

As a relatively robust form of prediction we compared the direction of extrapolated Markov chains for selected sites by ordination. For one site, trajectories of extrapolated Markov chains of plots showed a pattern typical for a general trend. In another case, trajectories indicated the absence of trend or the existence of site-dependent trends.

The comparison of the two methods by simulation revealed two trend qualities that are not necessarily correlated, namely the rate of change and its directedness. While the Markovian method accurately separated trend and noise, but did not distinguish between fast or slow change, the multivariate approach reacted reversely. The independence of the two qualities was further stressed by the fact that the respective scores of the 47 plots of the dry meadow data were not significantly correlated. Through the use of the two complementary methods, it has been possible both to detect and prove even slow adaptation processes that were obscured by much noise and to interpret their relevance in terms of the process rate.

To avoid misinterpretation, change should always be examined in terms of the two qualities, trend and rate. Markovian models are a powerful tool to assess trend in vegetation change. To make their whole potential available for prediction, alternative sampling strategies or transformations of the data may have to be evaluated.

Acknowledgements: This project has been supported by the Department of Interior of the Canton Aargau. The authors express their thanks to Dr. R. Maurer and A. Stapfer, Aarau, who encouraged this investigation and provided the data. We profited from suggestions by D. R. Cox, Nuffield College, Oxford, United Kingdom, and M. Anand, University of Western Ontario, London, Canada.

Résumé: Comment les chaînes de Markov peuvent-elles, en comparaison avec d'autres méthodes d'analyse multivariée plus traditionnelles, contribuer à l'analyse des tendances de courtes séries temporelles à plusieurs variables et présentant de fortes fluctuations? Deux méthodes ont été comparées à l'aide de la simulation et de l'analyse de données réelles d'un programme de monitoring de prairies sèches dans le canton d'Argovie en Suisse. La méthode markovienne distingue avec précision les tendances et les bruits simulés tandis que l'approche basée sur la régression ne mesure que le taux de tendance. La combinaison des deux méthodes au niveau de la parcelle et de la station a montré une tendance générale dans cinq stations, une prédominance des bruits dans une, un développement divergent sur différentes parts des prairies dans deux et des résultats ambigus dans une station. Pour éviter une fausse interprétation, le développement de-

vrait toujours être examiné en termes de taux et de tendance. Les modèles markoviens sont un précieux outil pour évaluer la tendance des changements phytosociologiques. Afin d'utiliser pleinement le potentiel d'extrapolation, il sera nécessaire d'étudier l'effet produit sur le modèle par divers modes d'échantillonnage et diverses transformations.

Mots clés: Bruit, données multivariées, prairies sèches, séries temporelles, tendance.

References

- [1] Austin, M. P. (1977), "Use of Ordination and Other Multivariate Descriptive Methods to Study Succession", *Vegetatio*, 35, 165-175.
- [2] Ball, G. H. (1966), *A Comparison of Some Cluster Seeking Techniques*, Stanford Research Institute, California.
- [3] Fekete, G. (1985), "Problems of Coenological Succession", *Magyar*, 31-63.
- [4] Gauch, H. G. (1982), *Multivariate Analysis in Community Ecology*, Cambridge University Press, Cambridge.
- [5] Jaccard, P. (1912), "The Distribution of the Flora of the Alpine Zone", *New Phytologist*, 11, 37-50.
- [6] Jongman, R. H. G., ter Braak, C. J. F., and van Tongeren, O. F. R. (1995), *Data Analysis in Community and Landscape Ecology*, Academic Press, Cambridge.
- [7] Lagonegro, M. (1984), "SPAGHET: A Coenocline Simulator Useful to Calibrate Software Detectors", *Studia Geobotanica*, 4, 63-99.
- [8] Maarel, E. van der, and Werger, M. J. A. (1978), "On the Treatment of Succession Data", *Phytocoenosis*, 7, 257-278.
- [9] Orlóci, L., Anand M., and He, X. (1993), "Markov Chain: A Realistic Model for Temporal Coenosere?", *Biom. Praxim.*, 33, 7-26.
- [10] Usher, M. B. (1981), "Modelling Ecological Succession, With Particular Reference to Markovian Models", *Vegetatio*, 46, 11-18.
- [11] Wildi, O. (1986), *Analysis of Vegetation Data*, Theory and Application of Numerical Methods, Veröff. Geobot. Inst. ETH, Stiftung Rübel, Zurich.
- [12] Wildi, O. (1988), "Linear Trend and Noise in Multi-Species Time Series", *Vegetatio*, 77, 51-56.
- [13] Wildi, O. (1990), "A Multiple Scale Sampling Design For Long Term Monitoring", Global Natural Ressource Monitoring and Assessments Preparing for the 21st Century, *Proc. of the International Conference and Workshop*, Bethesda USA, 975-982.

Contour Lines of L_1 -Norm Regression

Jean-Pierre Renfer

Statistics Group, University of Neuchâtel, Switzerland

Abstract: In linear regression, the parameters can be estimated by using different criteria. One of the most widely used criterion is the minimization of the sum of squared errors. Here we concentrate on another criterion, namely the minimization of the sum of absolute errors. This method is called L_1 -norm regression and is known for its robustness in presence of outliers. In the case of a simple linear model, the objective function to minimize will be studied by drawing its contour lines. The problem of constructing these contour lines is solved in the parameter plane. A simple algorithm is given to draw such contour lines.

Key words: Contour lines, linear regression, L_1 -regression.

1 Introduction

Consider the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n. \quad (1)$$

where y_i and x_i are the values of the dependent and independent variable respectively, ε_i an error term and β_0, β_1 the parameters to be estimated. The error term in (1) can be expressed as

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i, \quad i = 1, \dots, n. \quad (2)$$

Finding β_0 and β_1 from $(x_i; y_i)$ is called estimation of the parameters. Many different methods are available for obtaining such estimates. These methods are generally based on the minimization of some function of the residuals in (2). An important class of estimators are the L_p -estimators, defined by considering what is known as the Minkowsky metric or L_p -norm. These estimators are the solutions to the following minimization problem

$$\left(\sum_{i=1}^n |\varepsilon_i|^p \right)^{1/p} \text{ with } p \geq 1. \quad (3)$$

With this terminology, the least squares estimator is the L_2 estimator. For the L_1 -estimator, the function to minimize is

$$f(\beta_0, \beta_1) = \sum_{i=1}^n |y_i - \beta_0 - \beta_1 x_i| \quad (4)$$

The minimization of $f(\beta_0, \beta_1)$ in (4) is not trivial and the usual method using differential calculus cannot be used in this case. This is due to the fact that this function is not everywhere differentiable. The solutions are generally obtained by descent algorithms or by linear programming methods. For more details, see e.g. Birkes and Dodge (1993).

The goal in this paper is to give a new algorithm to construct exactly the contour lines of $f(\beta_0, \beta_1)$ without using an approximation method like in Snyder (1978) or Nakamura (1996). The mathematical development for achieving this goal is given in section 2. This development leads to an algorithm given in section 3. Finally, this algorithm is applied on a numerical example in section 4.

2 The contour lines

The contour lines of $f(\beta_0, \beta_1)$ are obtained by solving the equation

$$\sum_{i=1}^n |y_i - \beta_0 - \beta_1 x_i| = \sum_{i=1}^n f_i(\beta_0, \beta_1) = c \quad (5)$$

where $f_i(\beta_0, \beta_1) = |y_i - \beta_0 - \beta_1 x_i|$ and c such that

$$c \geq \sum_{i=1}^n |y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i|,$$

with c a constant and $\hat{\beta}_0$ and $\hat{\beta}_1$ the L_1 -norm estimates. The function $f_i(\beta_0, \beta_1) = |y_i - \beta_0 - \beta_1 x_i|$ can be defined as

$$f_i(\beta_0, \beta_1) = \begin{cases} y_i - \beta_0 - \beta_1 x_i & \text{if } \beta_0 \leq y_i - \beta_1 x_i \\ -y_i + \beta_0 + \beta_1 x_i & \text{if } \beta_0 \geq y_i - \beta_1 x_i \end{cases}$$

For each i , $f_i(\beta_0, \beta_1)$ can be viewed as two half-planes whose intersection is a line corresponding to $f_i(\beta_0, \beta_1) = 0$ having for equation $\beta_0 = y_i - \beta_1 x_i$. Another way of expressing the function $f_i(\beta_0, \beta_1)$ is to use the sign of $(y_i - \beta_0 - \beta_1 x_i)$ defined as follows

$$\text{sgn}(y_i - \beta_0 - \beta_1 x_i) = \begin{cases} +1 & \text{if } \beta_0 < y_i - \beta_1 x_i \\ 0 & \text{if } \beta_0 = y_i - \beta_1 x_i \\ -1 & \text{if } \beta_0 > y_i - \beta_1 x_i \end{cases} \quad (6)$$

which implies that

$$f_i(\beta_0, \beta_1) = (y_i - \beta_0 - \beta_1 x_i) \text{sgn}(y_i - \beta_0 - \beta_1 x_i).$$

The line segments forming the contour can be expressed using (5) and (6) to get $\beta_0 = a + b\beta_1$, where

$$a = \frac{-c + \sum_{i=1}^n y_i \text{sgn}(y_i - \beta_0 - \beta_1 x_i)}{\sum_{i=1}^n \text{sgn}(y_i - \beta_0 - \beta_1 x_i)} \quad (7)$$

$$b = -\frac{\sum_{i=1}^n x_i \text{sgn}(y_i - \beta_0 - \beta_1 x_i)}{\sum_{i=1}^n \text{sgn}(y_i - \beta_0 - \beta_1 x_i)} \quad (8)$$

If $\sum_{i=1}^n \text{sgn}(y_i - \beta_0 - \beta_1 x_i) = 0$, equations (7) and (8) do no longer apply. In that case, the segments of line are given by

$$\beta_1 = \frac{-c + \sum_{i=1}^n y_i \text{sgn}(y_i - \beta_0 - \beta_1 x_i)}{\sum_{i=1}^n x_i \text{sgn}(y_i - \beta_0 - \beta_1 x_i)} \quad (9)$$

Equations (7) and (8), respectively (9), do not change unless the sign of $(y_i - \beta_0 - \beta_1 x_i)$ is modified. So these equations hold only in the region of the plane for which $\text{sgn}(y_i - \beta_0 - \beta_1 x_i)$ is constant. A new segment of line is obtained when the sign of $(y_i - \beta_0 - \beta_1 x_i)$ changes. This happens when this segment intersects one of the lines $\beta_0 = y_i - \beta_1 x_i$, let say line k . At this intersection point, the sign of $(y_k - \beta_0 - \beta_1 x_k)$ is inverted and one can easily compute the equation of the next line segment. A contour line is completed by computing successively these equations. Before giving an algorithm to construct a contour line, let us clarify some notations. The n lines $\beta_0 = y_i - \beta_1 x_i$ divide the parameter plane in different regions according to their respective sign. For the points $(\beta_1; \beta_0)$ above (below) one of these lines, the sign is negative (positive). For example, the region for which $\text{sgn}(y_i - \beta_0 - \beta_1 x_i) = +1$ for $i = 1, \dots, n$ is denoted by $R_{+++...+}$. If

the first sign is negative and all other signs are positive, the corresponding region is $R_{-++\dots+}$ and so on. This notation is important because it gives a correspondence between a region and the values of $\text{sgn}(y_i - \beta_0 - \beta_1 x_i)$ which are necessary for calculating the segments of lines forming a contour line. Moreover, the value of c can be determined by choosing a point by which the contour line passes through.

We can now give an algorithm to construct a contour line passing through a selected point in the (β_1, β_0) plane.

3 Algorithm

This algorithm is based on the construction defined in section 2. It consists of the following steps.

Step 1 Draw the n lines $\beta_0 = y_i - \beta_1 x_i$ in the (β_1, β_0) plane.

Step 2 Choose the coordinates of a point $(\tilde{\beta}_0, \tilde{\beta}_1)$ through which the contour line will pass. Without loss of generality, this point can be chosen in the region $R_{+++ \dots +}$.

Step 3 Compute the value of c corresponding to this point by evaluating

$$c = f(\tilde{\beta}_0, \tilde{\beta}_1) = \sum_{i=1}^n |y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i|.$$

Step 4 In the current region compute from (7) and (8), respectively (9) the equation of the segment of line and draw it in the (β_1, β_0) plane until it intersects one of the line $\beta_0 = y_i - \beta_1 x_i$, say line k .

Step 5 Determine the new region by inverting the sign corresponding to the k th position.

Step 6 If the current region is the same as the first one, stop. Otherwise return to step 4.

Before giving an illustrative example of this algorithm, some remarks are needed. In Step 4, the current segment of line can intersect one of the line $\beta_0 = y_i - \beta_1 x_i$ on both sides. So one has to decide if the contour line will be constructed clockwise or anti-clockwise. It doesn't matter in which direction the contour line is constructed but once the direction has been chosen, it has to be kept through the algorithm. Note also that the current segment of line can intersect more than one line $\beta_0 = y_i - \beta_1 x_i$. This arises when the intersection point corresponds simultaneously to an intersection point of two (or more) lines $\beta_0 = y_i - \beta_1 x_i$. If the segment of line intersects at the same point line k and l , then in Step 5 the new region is determined by inverting both signs corresponding to the k th and l th position. The following section illustrates numerically this algorithm.

4 Example

Consider the following data from Brown (1980)

x_i	-2	-1	0	1	2
y_i	0	1	-2	-2	1

The objective function for this data set is

$$f(\beta_0, \beta_1) = |-\beta_0 + 2\beta_1| + |1 - \beta_0 + \beta_1| + |-2 - \beta_0| \\ + |-2 - \beta_0 - \beta_1| + |1 - \beta_0 - 2\beta_1|.$$

This function is represented in Figure 1.

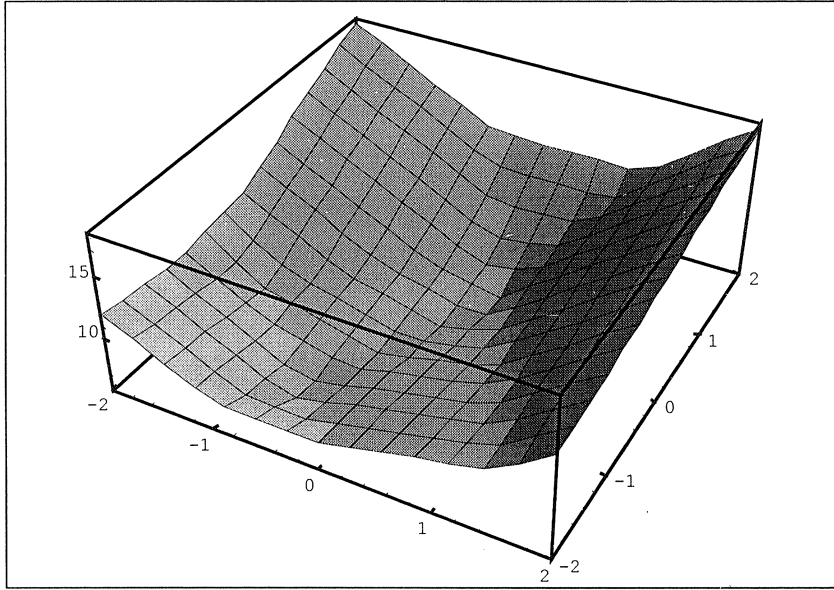


Figure 1: Graph of $f(\beta_0, \beta_1)$.

To construct a contour line, we apply the algorithm described in the preceding section.

Step 1 The five lines $\beta_0 = y_i - \beta_1 x_i$, ($i = 1, \dots, 5$) are drawn in Figure 2. They divide the (β_1, β_0) plane according to the sign of $(y_i - \beta_0 - \beta_1 x_i)$.

Step 2 Let us arbitrarily choose the coordinates of a point $(\tilde{\beta}_1; \tilde{\beta}_0) = (0; -2.6)$ in the region R_{+++++} , the region for which $\text{sgn}(y_i - \beta_0 - \beta_1 x_i) = +1$, $i = 1, \dots, 5$.

Step 3 We compute the value of c by evaluating

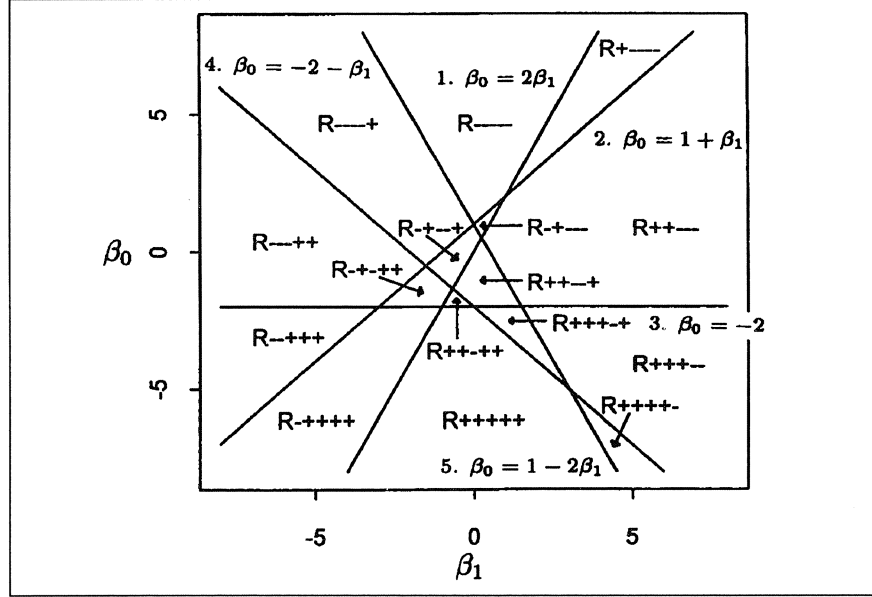


Figure 2: Lines $\beta_0 = y_i - \beta_1 x_i$ and the corresponding regions of the (β_1, β_0) plane.

$$f(\tilde{\beta}_0, \tilde{\beta}_1) = 2.6 + 3.6 + 0.6 + 0.6 + 3.6 = 11.$$

Step 4 To obtain the equation of the segment of the contour line in this region, one has to compute

$$\begin{aligned} \sum_{i=1}^5 \text{sgn}(y_i - \beta_0 - \beta_1 x_i) &= 5 \\ -c + \sum_{i=1}^5 y_i \text{sgn}(y_i - \beta_0 - \beta_1 x_i) &= -11 - 2 = -13 \\ \sum_{i=1}^5 x_i \text{sgn}(y_i - \beta_0 - \beta_1 x_i) &= 0 \end{aligned}$$

Using (7) and (8), the equation of the segment of line in this region is given by $\beta_0 = \frac{-13}{5} - \beta_1 \cdot 0 = -2.6$ and is drawn in the region R_{+++++} of Figure 3.

Step 5 The line segment obtained in Step 4 intersects line 4 at point $(0.6; -2.6)$. So $k = 4$. (One can also choose the other direction by using $k =$

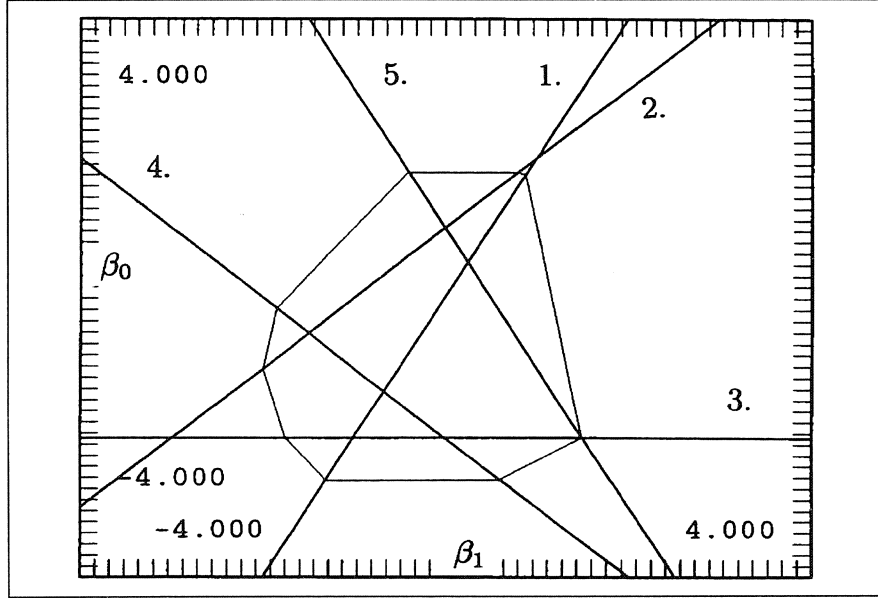


Figure 3: Contour line for $c=11$ in the (β_1, β_0) plane.

1 because the line segment also intersects line 1, the only difference is that the contour line will be completed clockwise rather than anti-clockwise).

Step 6 With $k = 4$, the new region in which equations (7) and (8) are valid is therefore R_{+++--} . So we return to Step 4.

To get the new equation of the line segment, one has now to compute the following expressions

$$\begin{aligned} \sum_{i=1}^5 \text{sgn}(y_i - \beta_0 - \beta_1 x_i) &= 3 \\ -c + \sum_{i=1}^5 y_i \text{sgn}(y_i - \beta_0 - \beta_1 x_i) &= -11 + 2 = -9 \\ \sum_{i=1}^5 x_i \text{sgn}(y_i - \beta_0 - \beta_1 x_i) &= -2 \end{aligned}$$

resulting in the following equation

$$\beta_0 = \frac{-9}{3} + \frac{2}{3}\beta_1 = -3 + \frac{2}{3}\beta_1.$$

This segment of the contour line is also drawn in Figure 3. It intersects both

Region R	Equation of the line segment	Intersection with line
+++++	$\beta_0 = -2.6$	4
+++ - +	$\beta_0 = -3 + \frac{2}{3}\beta_1$	3 and 5
++ - - -	$\beta_0 = 7 - 6\beta_1$	1
- + - - -	$\beta_0 = \frac{7}{3} - \frac{2}{3}\beta_1$	2
- - - - -	$\beta_0 = 1.8$	5
- - - - +	$\beta_0 = \frac{7}{3} + \frac{4}{3}\beta_1$	4
- - - + +	$\beta_0 = 11 + 6\beta_1$	2
- + - + +	$\beta_0 = -9 - 4\beta_1$	3
- + + + +	$\beta_0 = -\frac{13}{3} - \frac{4}{3}\beta_1$	1
+++++	$\beta_0 = -2.6$	

Table 1: Equations of the segments of line forming the contourline $f(\beta_0, \beta_1) = 11$.

lines 3 and 5. So the new region is $R_{++-----}$. To obtain the complete contour line, we iterate this process until the current region is again R_{+++++} .

Table 1 summarizes the iterations of the algorithm by indicating the regions, the corresponding equations and the index of the intersection line.

The contour line for $c = 11$ and the different regions of the plane are represented in Figure 3.

Note that for each region there exists a value of c for which the segment of line can no longer belongs to that region.

For example, in the region R_{+++++} , the segment of line has the following equation

$$\beta_0 = \frac{-c - 2}{5}.$$

Geometrically, in this region, these segments of line run parallel to each other according to the value of c . But one cannot move these segments of line further than the intersecting point between line 1 and 5. This intersection point is (1.5;-2) so that the limit value of c in that region is determined by

$$-2 = \frac{-c - 2}{5} \Rightarrow c = 8.$$

This particular region is not limited so that the condition is $8 \leq c < \infty$. For the limited regions (closed), the value of c lies between two finite values. A summary of these values of c for each region is given in Table 2.

The minimal possible value of c is 6 and corresponds to the sum of least absolute deviations. Knowing this minimal value of c , the contour lines

Region R	Equation of the line segment	Values of c
+++++	$\beta_0 = \frac{-c-2}{5}$	$8 \leq c < \infty$
++++-	$\beta_0 = \frac{-c-4}{3} + \frac{4}{3}\beta_1$	$23 \leq c < \infty$
+++ - +	$\beta_0 = \frac{-c+2}{3} + \frac{2}{3}\beta_1$	$8 \leq c \leq 23$
+++ --	$\beta_0 = -c + 6\beta_1$	$11 \leq c < \infty$
++ ---	$\beta_0 = c - 4 - 6\beta_1$	$6 \leq c < \infty$
++ -- +	$\beta_0 = -c + 6 + 2\beta_1$	$6 \leq c \leq 11$
+ ----	$\beta_0 = \frac{c-2}{3} - \frac{4}{3}\beta_1$	$12 \leq c < \infty$
- + ---	$\beta_0 = \frac{c-4}{3} - \frac{2}{3}\beta_1$	$6 \leq c \leq 12$
-----	$\beta_0 = \frac{c-2}{5}$	$7 \leq c < \infty$
- + - - +	$\beta_0 = c - 6 + 2\beta_1$	$6 \leq c \leq 8.5$
----- +	$\beta_0 = \frac{c-4}{3} + \frac{4}{3}\beta_1$	$7 \leq c < \infty$
--- ++	$\beta_0 = c + 6\beta_1$	$8.5 \leq c < \infty$
- + - + +	$\beta_0 = -c + 2 - 4\beta_1$	$6 \leq c \leq 16$
-- + + +	$\beta_0 = -c - 4 - 6\beta_1$	$16 \leq c < \infty$
+ + - + +	$\beta_0 = \frac{-c+2}{3}$	$6 \leq c \leq 8$
- + + + +	$\beta_0 = \frac{-c-2}{3} - \frac{4}{3}\beta_1$	$8 \leq c < \infty$

Table 2: Values of c for which the given equation applies in the corresponding region.

are easily computed by applying the algorithm from Step 4. This has been done in Figure 4 for different values of c .

Since the objective function is convex, the contour lines are also convex. The minimum is reached on the segment of line that connects the intersection points of line 1 with lines 4 and 5 respectively. This is a situation in which the L_1 -norm estimates are not unique.

Finally, in each region, the contour lines are parallel. This property can be used for drawing quickly the contour lines in the different regions (the maximal number of different regions is $1 + n(n + 1)/2$).

Résumé: En régression linéaire les paramètres peuvent être estimés selon différents critères. L'un des plus utilisé est le critère de la minimisation de la somme des carrés des erreurs. Nous nous concentrons ici sur un autre critère, la minimisation de la somme des valeurs absolues des erreurs. Cette méthode, appelée régression L_1 , est connue pour sa robustesse en présence de valeurs aberrantes. Dans le cas d'un modèle linéaire simple, la fonction objectif à minimiser est étudiée à travers ses courbes de niveaux. Le problème de la construction de ces courbes de niveaux est résolu dans le plan des paramètres. Un algorithme simple est donné pour dessiner ces

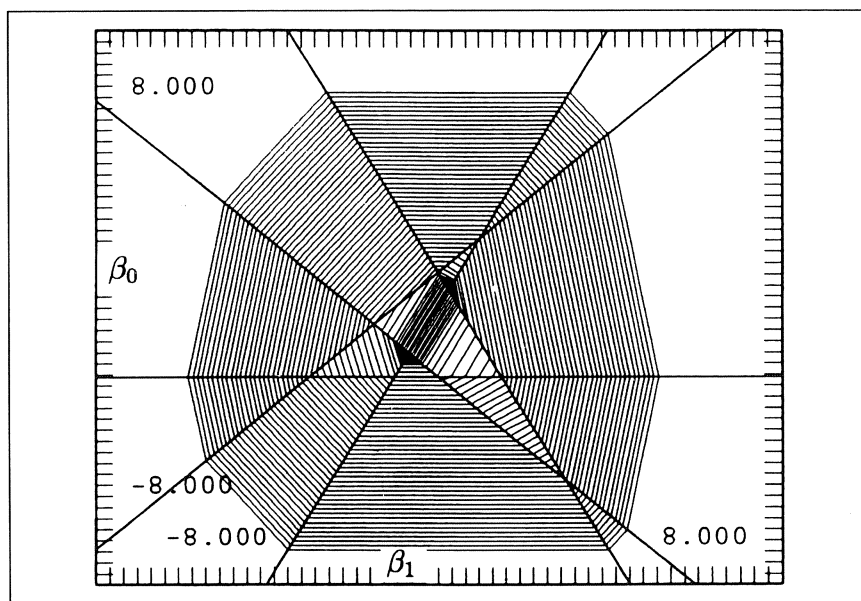


Figure 4: Contour lines for $6 \leq c \leq 33$.

courbes de niveaux.

Mots clés: Courbes de niveaux, régression linéaire, régression L_1 .

References

- [1] Birkes, D. and Dodge, Y. (1993), *Alternative Methods of Regression*, John Wiley, New York.
- [2] Brown, B.M. (1980), "Median Estimates in Simple Linear Regression", *Austral. J. of Statist.*, 22, 154-165.
- [3] Nakamura, S. (1996), *Numerical Analysis and Graphic Visualization with Matlab*, Prentice Hall Inc., New Jersey.
- [4] Snyder, W.V. (1978), "Contour Plotting [J6]", *ACM Transactions on Mathematical Software*, 4, 290-294.

From the Editor's Notebook

Visualizing R^2

Let X and Y be two random variables related as

$$Y = a + bX + \varepsilon$$

where ε is a random variable independent from X . As

$$Cov(X, Y) = Cov(X, bX) + Cov(X, \varepsilon) = b \cdot Var(X) + 0,$$

it follows that

$$b = \frac{Cov(X, Y)}{Var(X)}.$$

Given a sample of n bivariate observations (x_i, y_i) of (X, Y) , the usual estimator of b is then

$$\hat{b} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \text{ where } \bar{x} = \frac{\sum x_i}{n} \text{ and } \bar{y} = \frac{\sum y_i}{n}.$$

The coefficient of determination R^2 which describes the strength of the linear association between X and Y , is defined as

$$R^2 = 1 - \frac{Var(\varepsilon)}{Var(Y)} = 1 - \frac{Var(Y) - b^2 \cdot Var(X)}{Var(Y)} = b^2 \cdot \frac{Var(X)}{Var(Y)}.$$

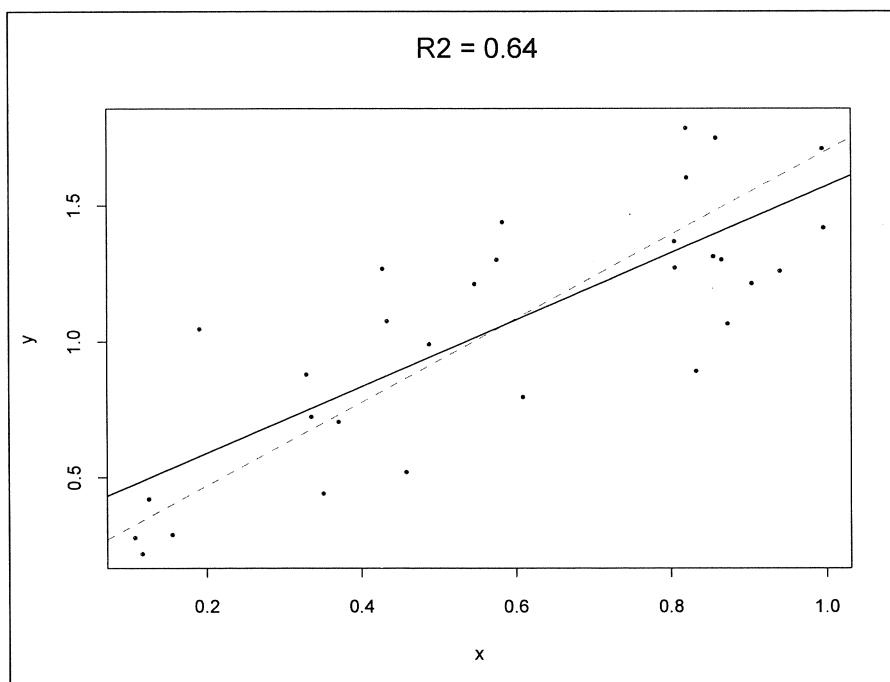
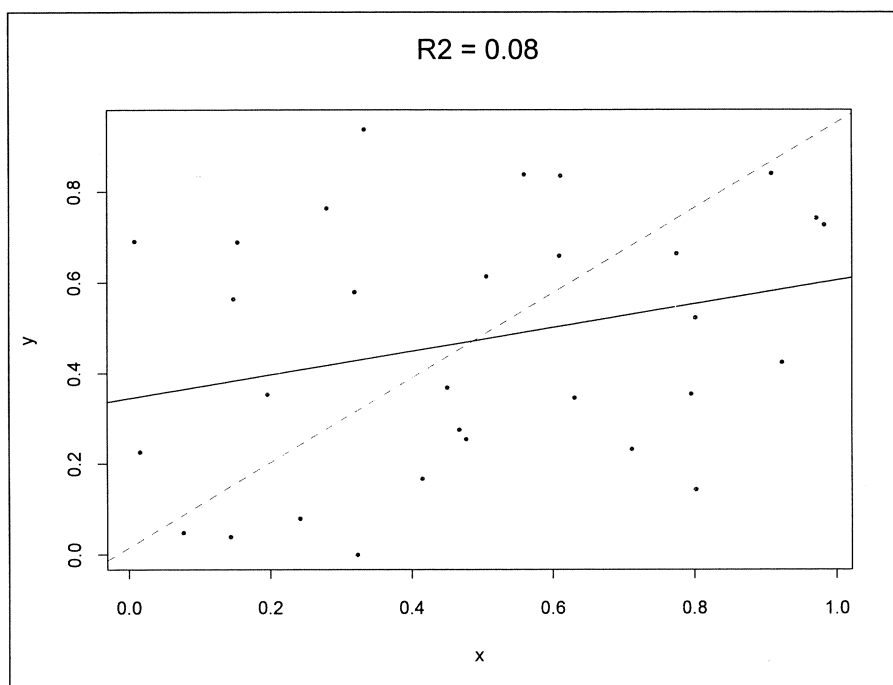
The maximal value of R^2 is 1 and is attained when $Var(\varepsilon) = 0$, or equivalently when

$$b^2 = \frac{Var(Y)}{Var(X)}.$$

In order to illustrate graphically R^2 , we propose to plot on a same graph a solid line with slope $\hat{b}$, and a dashed line with slope

$$\tilde{b} = \text{sign}(\hat{b}) \cdot \sqrt{\frac{\sum (y_i - \bar{y})^2}{\sum (x_i - \bar{x})^2}}$$

which intersect at the point $(\bar{x}, \bar{y})$. The greater R^2 is, the closer $\hat{b}$ and $\tilde{b}$ are. The two lines coincide exactly when $R^2 = 1$. The two figures on the next page illustrate this kind of graph with 30 observations where respectively $R^2 = 0.08$ and $R^2 = 0.64$.



Irving Fisher: “I rather work to death than rust to death idly”



Irving Fisher was born on February 27, 1867, in Saugerties, New York. His skills in using mathematics in economic science justify calling him the first mathematical economist of America. Throughout his career he made durable contributions not only in economic theory but also in statistical inquiry. Fisher's doctoral dissertation (1892) at Yale University won him worldwide recognition in economic theory, especially in equilibrium theory. He founded the Econometric Society in 1930 together with Ragnar Frisch and Charles F. Roos, and became its first President after having been President of the American Economic Association in 1918.

Fischer stayed at Yale throughout his career. He started teaching mathematics before he received his doctorate. His first economics teaching was under the auspices of the mathematics faculty, a course on "The mathematical theory of prices". In 1894-95 he visited several Universities in Europe and was welcomed by the mathematically inclined theorists in every country. When he returned to America, he became Assistant Professor of Political and Social Science. He was appointed full Professor in 1898 and retired in 1935. He died in New York on April 29, 1947. His semicentennial year of death is taking place this year on April 29.

One can cite from Irving Fisher (1947):

“From the beginning and throughout my life ‘I have tried’ to help build economics into a genuine science, comparable with the physical sciences, as well as to apply its principles to help solve certain practical economic problems pressing for solution”

Fisher was a very prolific writer with more than 2000 titles authored as compiled in the bibliography written by his son in 1961. He was the major American monetary economist of the early decades of this century and combined theorizing with empirical research, both historical and statistical. He encountered problems that led him to invent statistical and econometric methods, in particular index numbers. In 1920 he proposed as the “Ideal Index” a candidate, namely the geometric mean of the Laspeyres and Paasche formulas, having the pleasant property that the correlate of an Ideal price index is an Ideal quantity index.

Selected publications of Irving Fisher are:

- (1892) *Mathematical Investigations in the Theory of Value and Prices*, Connecticut Academy of Arts and Science, New Haven.
- (1906) *The Nature of Capital and Income*, Macmillan, New York.
- (1907) *The Rate of Interest*, Macmillan, New York.
- (1910) *Introduction to Economic Science*, Macmillan, New York.
- (1911) *The Purchasing Power of Money*, Macmillan, New York.
- (1911) *Elementary Principles of Economics*, Macmillan, New York.
- (1921) “The Best Form of Index Number”, *American Statistical Association Quarterly*, 17, 533-537.
- (1922) *The Making of Index Numbers*, Houghton Mifflin, Boston.
- (1926) “A Statistical Relation Between Unemployment and Price Changes”, *International Labour Review*, 13, 785-792.
- (1927) “A Statistical Method for Measuring ‘Marginal Utility’ and Testing the Justice of a Progressive Income Tax”, in *Economic Essays Contributed in Honor of John Bates Clark*, ed. J.H. Hollander, Macmillan, New York.
- (1930) *The Theory of Interest*, Macmillan, New York.
- (1932) *Booms and Depressions*, Adelphi, New York.
- (1933) “The Debt-Deflation Theory of Great Depressions”, *Econometrica*, 1, 337-357.
- (1935) *100% Money*, Adelphi, New York.

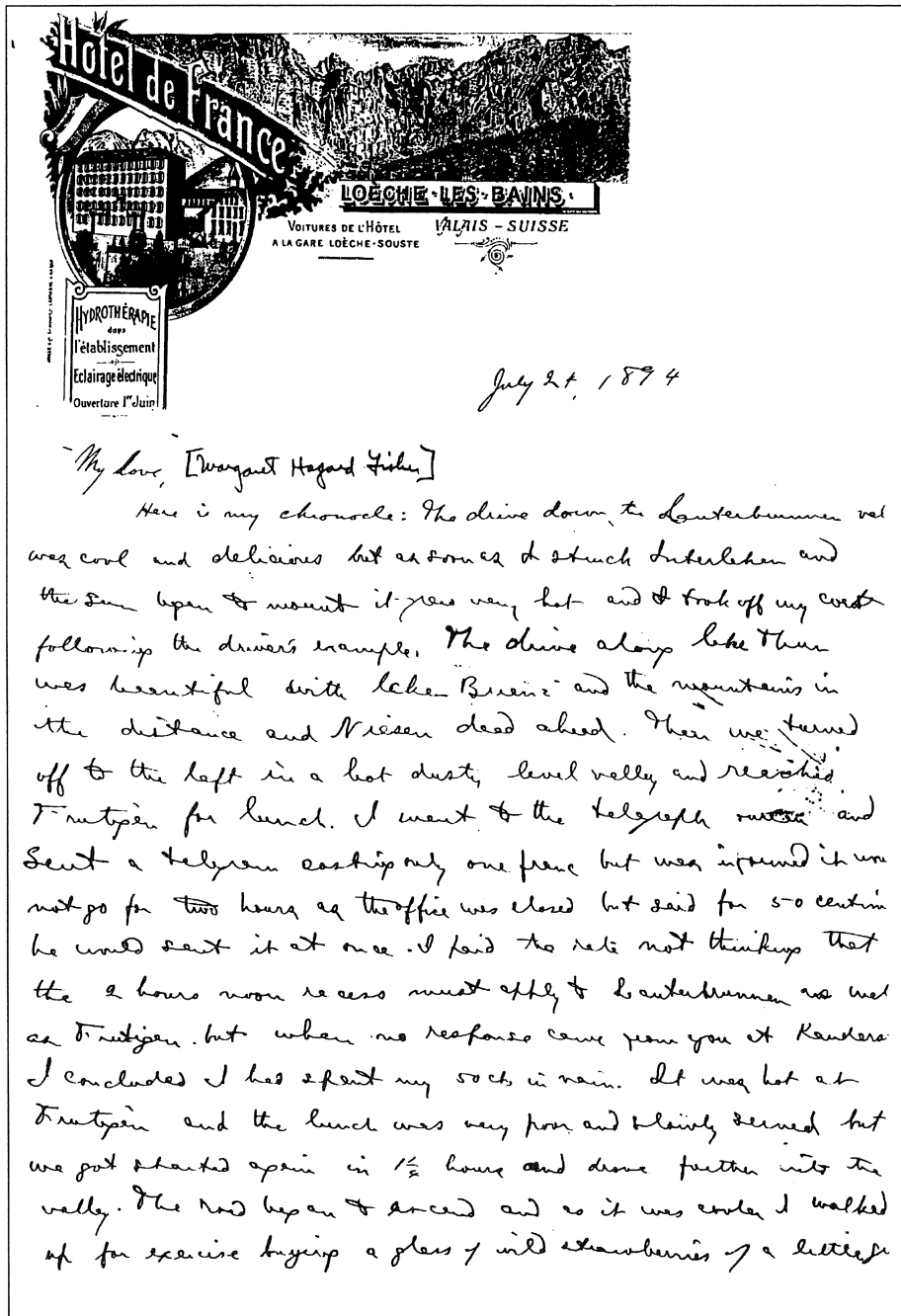
In the honour of the 50th anniversary of his death, A.Vogt and J. Barta recently published a book entitled *The Making of Tests for Index Numbers. Mathematical Methods of Descriptive Statistics*.

The secret of Fisher's working power was best expressed by Cohrssen of Frankfurt who worked from 1932 to 1942 for Fisher in a talk given at the occasion of Fisher's 80th birthday:

"It was in 1896 in a tuberculosis sanatorium, high in the Mountains of Arizona, where the air is pure and clear. A young man of 29 was sitting in his invalid-chair waiting for death. The doctors had given him only a very short time to live. Behind him were years of the most intensive work as an economist at a university and exceptional academic recognition. This man, marked by death, looked at his young wife, his child playing before him in the sand.... and he made up his mind: "I rather work to death than rust to death idly." He left the mountains and returned to his work... He took up the battle against his sickness and packed into each and every day so much work that in the race against his death he might leave at least the maximum of creative work of which he was capable. Week after week, month after month, year after year this feverish race continued... One book after another appeared, basic contributions to the science of economics, also notes recording his experiences in the fields of health and hygiene which he made in quest for life extension... He has been adviser to several presidents [of USA]; he was one of the founders of the Peace Society from which later developed the League of Nations... Today he is 80. He is still tirelessly active and one of the happiest men I've ever met. His name is Irving Fisher, Professor of Yale University..."



We provide the readers of *Student* with a copy of the first page from a six page letter that Irving Fisher wrote to his wife on July 24, 1894 describing a tour he made from Lauterbrunnen to Zermatt, Switzerland.



Fisher's letter to his wife from Switzerland. Reprinted by permission from Manuscripts and Archives, Yale University Library.

Chikio Hayashi and Data Science



Chikio Hayashi was born on June 7, 1918, in Tokyo, Japan. He received a B.A. in Mathematics from the University of Tokyo in 1942, and a D.Sc. from the same University in 1954. From 1955 he was a Professor at the Institute of Statistical Mathematics (National Research Institute) where he became Director-General in 1974 and Professor Emeritus in 1986. From 1986 to 1991 he was Professor at the University of the Air. Since then he is Visiting Professor in 12 Japanese universities. In 1965, Chikio Hayashi received the NHK Broadcasting Culture Prize for his research in social survey and in theory of quantification. In 1977, he was named Honorary Fellow of the Royal Statistical Society. In 1982 he has been awarded the Purple Ribbon Medal from the Japanese Government.

Doctor Hayashi is the author or coauthor of 28 books in Japanese language in different fields of Statistics. He especially wrote books which treat on the Japanese National Character such as *Measurement of Japanese Mind* (1988) and *Structure of Japaneseness* (1996). He is also the author of books in English such as *Data Analysis for Comparative Social Research* (1992), *Treatise on Behaviormetrics* (1993) and *Quantification of Qualitative Data - Theory and Method* (1993).

Chikio Hayashi was the student of Professor Z. Suetuna of the Department of Mathematics of the Tokyo University. He was a senior colleague of two other famous Japanese statisticians G. Taguchi and H. Akaike who had the same teacher as him.

G. Taguchi is famous for his method for off-line quality control which brought a lot of admiration from practical industry including the United States (Taguchi, G. (1976), *An Introduction to Quality Control*, Central Japan Quality Control Association, Nagoya, Japan).

Hirotsugu Akaike is the founder of a criterion which is used to decide on the order of a regression, where there is a natural sequence for introduction of successive predictor values, for example, ARIMA. An Information Criterion is defined by the relation

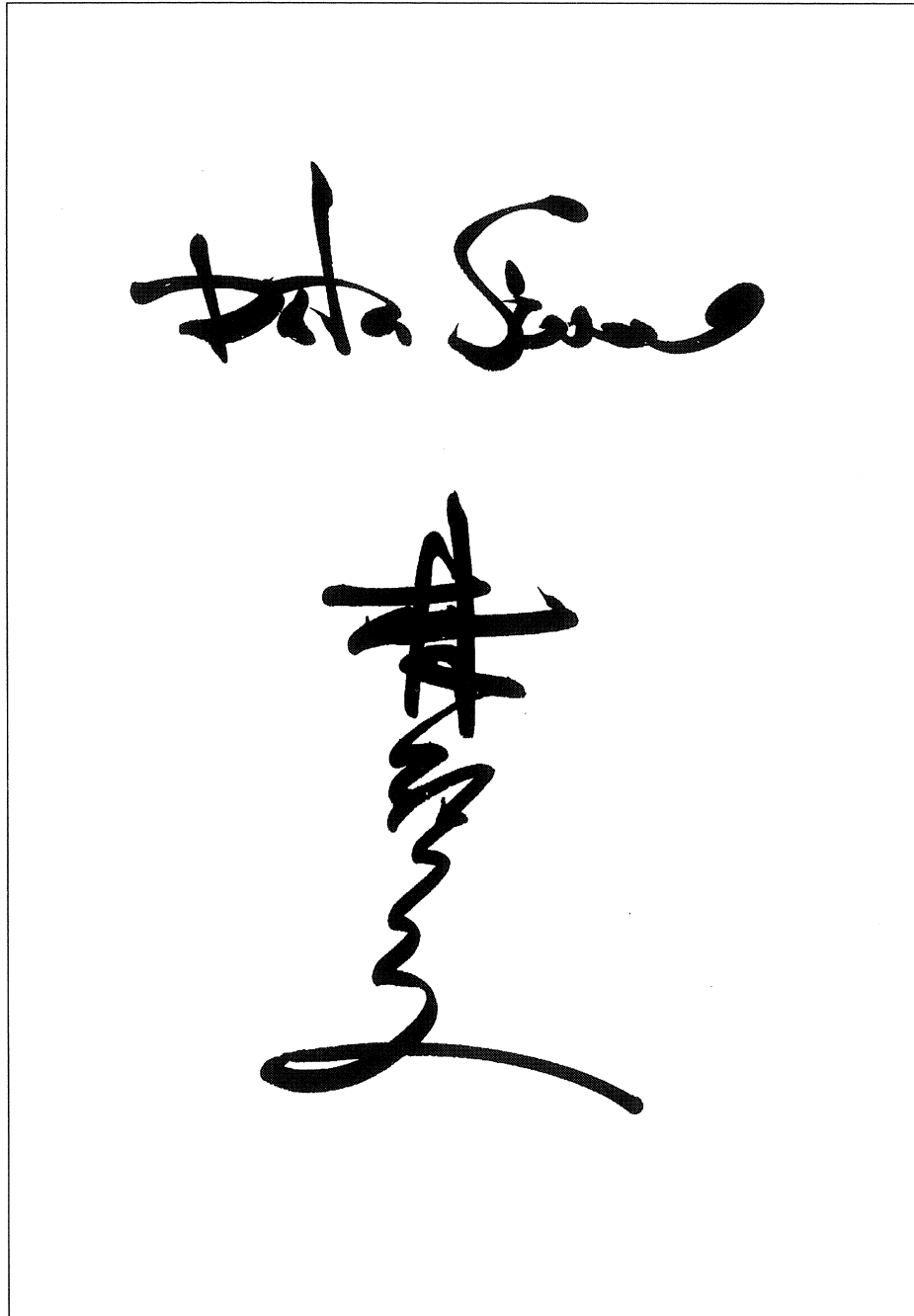
$$AIC = -2 \cdot \text{maximum log likelihood} + 2 \cdot \text{number of parameters}$$

(Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle" in *2nd International Symposium on Information Theory*, B.N. Petrov and F. Csaki eds., Akademiai Kiado, Budapest).

In what follows we provide the reader with Professor Hayashi's point of view on what is *Data Science* and how it is developed in Japan, a country for which we may not be familiar with.



Chikio Hayashi giving lecture during the IFC 1996 Conference, Kobe, Japan.



Chikio Hayashi writes on *Data Science*.

What is Data Science ?

- Fundamental concept -

By Chikio Hayashi

Institute of Statistical Mathematics

1. Introduction

Statistics and data analysis have developed in their realms separately. They contributed to the development of science, showing their properties. The idea and various methods of statistics were very useful as well-known and solved many problems. Mathematical statistics succeeded to it

and improved new frontiers by the idea of statistical inference. Thus the elevation of view point brought us many useful results. On the other hand, the method of data analysis has developed in the fields disregarded by mathematical statistics and given useful results to solve complicated problems based on mathematico-statistical methods (which are not always on statistical inference but rather descriptive). As the differentiation proceeds with specialization, the innovation of useful methods of statistics and data analysis seems to disappear and a sign of stagnation appears. The reason is that the essential aim of analysis of phenomena by data has been forgotten. For extensive and profound development of intrinsically useful methods of statistics and data analysis beyond present state, the necessity of unification of statistics and data analysis is felt acutely. For this purpose, the construction of a new point of view or a new concept is a crucial problem. So, I will present *Data Science* as a new concept.

2 Fundamental concept of data science

Data science is not only a synthetic concept to unify statistics, data analysis and their related methods on data, but also comprises its results. Data science intends to analyze and understand actual phenomena by *data*. In another expression, the aim of data science is to reveal the features or the

hidden structure of complicated natural, human and social phenomena by data from the different point of view from the established or traditional theory and method. This point of view implies multidimensional, dynamic and flexible ways of thinking.

Data Science consists of three phases: *design for data*, *collection of data* and *analysis on data*. It is important that three phases are treated in the concept of unification based on the fundamental philosophy of science explained as below. At these phases, the methods, which are fitted for the object and valid, must be studied in a good perspective. The strategy for research in data science through three phases is summarized in Figure 1.

Generally speaking, phenomena are multifarious. First, these phenomena are formulated and the planning of survey or experiment is devised, based on the idea of data science (phase of design for data). Thus phenomena are expressed as multidimensional and, frequently, time-series data. The characteristics or properties of these data are necessarily made clear (phase of collection of data). The obtained data are too complicated to draw clear conclusion. So, by methods of classification and multidimensional data analysis, and other mathematico-statistical methods, the data structure is revealed, in other words, simplification and conceptualization are carried out. However, this information generally turns out to be incomplete and unsatisfactory even though the structure finding was realized. At this stage, by finding and reconsidering the deviation of “individuals”, which gives vivid account of roughness of conceptualization or simplification, from the mean values or class-belonging (classification) and structure, diversification of data is made. Based on this multifariousness, structure finding or conceptualization is attained, in an advanced sense, on the progressive stage. Such a circular movement of research always continues. Dynamics of both simplification or conceptualization and diversification begin in turn. Further, having been able to solve a problem, it is expected to find out another new problem to be solved in an advanced sense. The developmental process,

$$\begin{aligned} &\text{design} \rightarrow \text{collection} \rightarrow \text{analysis} \rightarrow \text{design} \rightarrow \text{collection} \rightarrow \text{analysis} \rightarrow \\ &\text{design} \rightarrow \text{collection} \rightarrow \text{analysis} \rightarrow \text{design} \rightarrow \dots \end{aligned}$$

in phase, and dynamic process mentioned above, that is to say, progress and regress are indispensable in data science. It is expressed that methodology of data science develops, as it were, in the ascending-spiral-process and research proceeds as seen in spiral stairs. The main point is schematically depicted in Figure 1.

Thus we can say that data science comprises not only the results themselves of theory and method but also all methodological results related to

various processes which are necessary to work out the results mentioned above. The former is called *hard results* and the latter is called *soft results*. Data science includes simultaneously hard and soft results. It goes without saying that a useful solution emerges in coping with the complicated problem in question by data science. It is repeatedly emphasized that the coherent idea through all items shown in Figure 1 flows in data science for the purpose of analysis of phenomena by data.

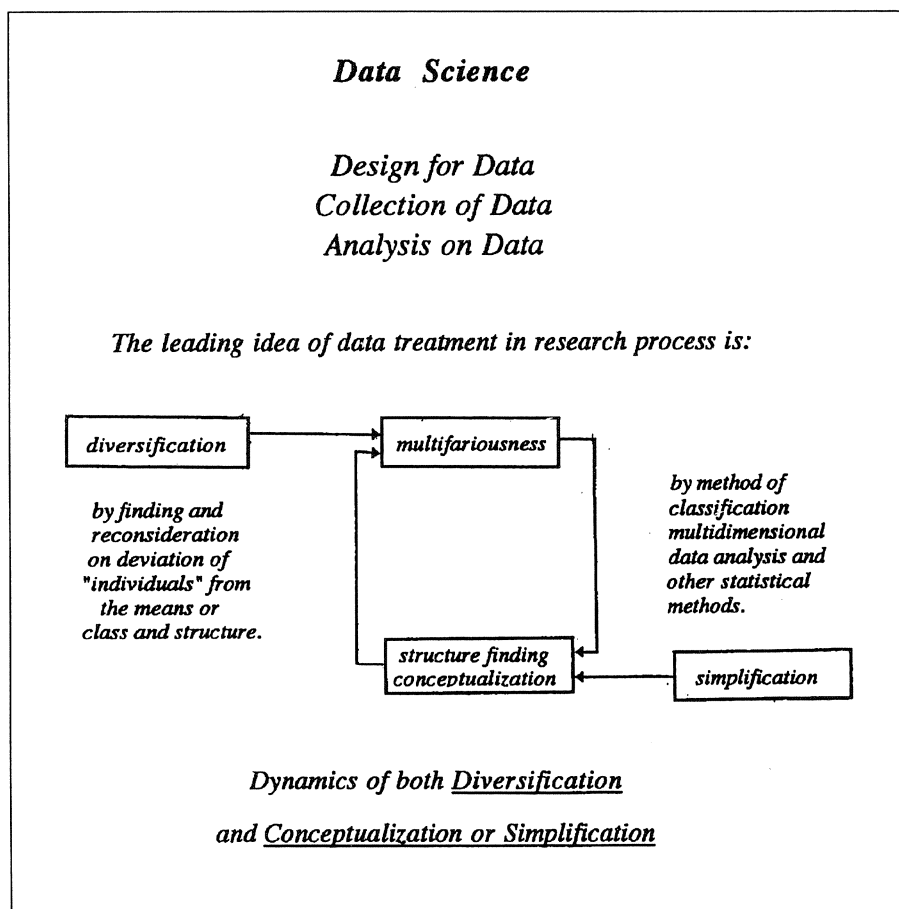


Figure 1: Strategy for research.

3 Content of data science

Some concrete examples in social and medical surveys for three phases are shown as below. Before everything, it is stressed that the relevant methods are always treated with validity.

3.1 How to design

The theory and method concerning this phase are considered. Particularly, theoretical and systematic construction of a questionnaire is a very important problem with the methods of observation and experiment. The problems in this phase are frequently solved using various kinds of methods of data analysis. For example,

- Sampling survey methods,
- Design of experiment,
- Evaluation of bias in quota sampling,
- New systematic idea of survey planning for the solution of difficult and complicated problem,
- Construction of questionnaire,
 - Theory-driven (which is on the extension of hypothesis testing), Guttman's Facet Theory,
 - Data-driven (exploratory approach), Hayashi's cultural Link Analysis in comparative study,
- Device of various types of questions, for example, dynamic use of closed and open ended questions,
- Use of various projective methods,
- Evaluation of data quality and data characteristics,
- Randomized response method,
- Problem of translation in international comparative study, and etc.

3.2 How to collect

Collection of data is not only a problem of practice, but must be theoretically and concretely studied. The problems in this phase can not be solved without any information of design for data and any use of data analysis.

- Evaluation of survey bias and evaluation of experimental bias including question bias, interview bias, interviewer bias, observation bias and etc.,
- Evaluation of non-response error,
- Evaluation of measurement error,
- Evaluation of response error, inevitably variable response data, for example, live data,
- Method of diminution of the relevant bias and error, and etc.

3.3 How to analyze data

The problems in this phase are, of course, closely related to the previous two phases. The main point is to obtain useful and instrumental information without any distortion or with validity. For this purpose, clear and lucid methods of analysis without unnecessary mathematical conditions only for

model building and too sophisticated style are desirable. For example,

- Various methods of scaling, quantification methods, correspondence analysis (analyse des données), multidimensional scaling, exploratory data analysis, categorical data analysis and various methods of classification and clustering,

- Useful data analysis suitable for the purpose,
- Useful coding of questions and their synthesis,
- Valid analysis of data including various errors,
- Evaluation of data quality and data analysis depending on data quality,
- Analysis on probabilistic response,
- Exploratory approach by data analysis,
- Method of simultaneous realization of classification and structure find-

ing,

- Treatment of open answers in an open ended question, for example, exploratory approach for coding or automatic processing of textual data,
- Probabilistic approach,
- Computer experiments, and etc.

These three phase must be synthetically treated or taken into consideration on the consistent idea in order to understand phenomena by data. This is the fundamental concept of data science. Of course, each subject will be studied separately. However, each subject must be studied in the context of data science. This idea will lead to development of statistics and data analysis in a new direction. Thus the stand points of them heighten and a new horizon will appear. Innovative method and theory will be created in three phases.

Chih Flya

*Tokyo, Japan
April, 1996.*

***Myrmecia* Measurement Data**

Nagatomo Nakamura

Department of Elementary Education, Kyushu Women's Junior College, Japan

Noboru Ohsumi

The Institute of Statistical Mathematics, Japan

Keiichi Onoyama

Obihiro University of Agriculture and Veterinary Medicine, Japan

Kazuo Ogata

Institute of Tropical Agriculture, Kyushu University, Japan

Robert W. Taylor

CSIRO Division of Entomology, Australia

Abstract: The purpose of the present paper is to show raw data of measurements of the Australian ant genus *Myrmecia* for the further statistical analysis. Twelve measurement data obtained from 252 individuals belonging to 82 species of the genus *Myrmecia* are presented. The data are arranged following 9 species groups. Basic statistics including numbers of individuals measured and missing values, average, standard deviation, coefficient of variation, minimum and maximum values are summarized. Additionally, valid numbers of individuals of 10 ratio variables which will be useful for distinguishing species or species-groups is also given.

Key words: Australian ants, cluster analysis, measurement data, *Myrmecia*, taxonomy.

1 Background

The genus *Myrmecia* is the sole extant representative of the ant subfamily Myrmeciinae. The genus contains about 100 species, distributing in Australian Continent and its adjacent areas, including Tasmania, New Caledonia, and New Zealand. In their course of taxonomic study, Taylor prepared the materials, arranging specimens preserved in CSIRO collection or on loan from various institutions, and Ogata compared these specimens, measuring individuals and analyzing other morphological characters. The phylogenetic analysis of species groups was presented by Ogata (1991) and the taxonomic review of the species was published by Ogata and Taylor (1991).

In these papers, however, the measurement data were presented only in a briefly summarized form. In fact, Ogata collected measurement data from 252 specimens belonging to 82 species of *Myrmecia*. Based on the idea that the raw data of the measurements are valuable in terms of statistical reviews, Nakamura, Ohsumi, and Onoyama rearranged them for the purpose of further statistical analysis. The present paper gives the list of measurements of *Myrmecia* and a summary of basic statistics.

2 The measurements

The data set consists of 12 variables. These are head width (HW), head length (HL), scape length (SL), mandible length (ML), Weber's length (WL), pronotal width (PrW), hind femoral length (HFL), petiole width (PtW), petiole length (PtL), postpetiole width (PpW), postpetiole length (PpL) and gastral width (GW). The definitions of the measuring were given by Ogata and Taylor (1991), except for the hind femoral length (the maximum length of the hind femur). Figure 1 briefly shows the measuring portions, according to them. Since the measurements were carried out by a single person (Ogata alone), the error caused by different observers can be eliminated.

3 Tables and figure for the data

Table 1 lists nine species groups classified by using morphological characters (see Ogata, 1991). Figure 1 shows a part of ant's body for the measurements. Table 2 shows various indices; those of the upper portion are used by Ogata (1991) and Ogata and Taylor (1991); lower portion by Nakamura (1995a). Table 3 is basic statistics obtained from the raw measurement data.

The complete list of measurements is shown in Table 4. Dividing each

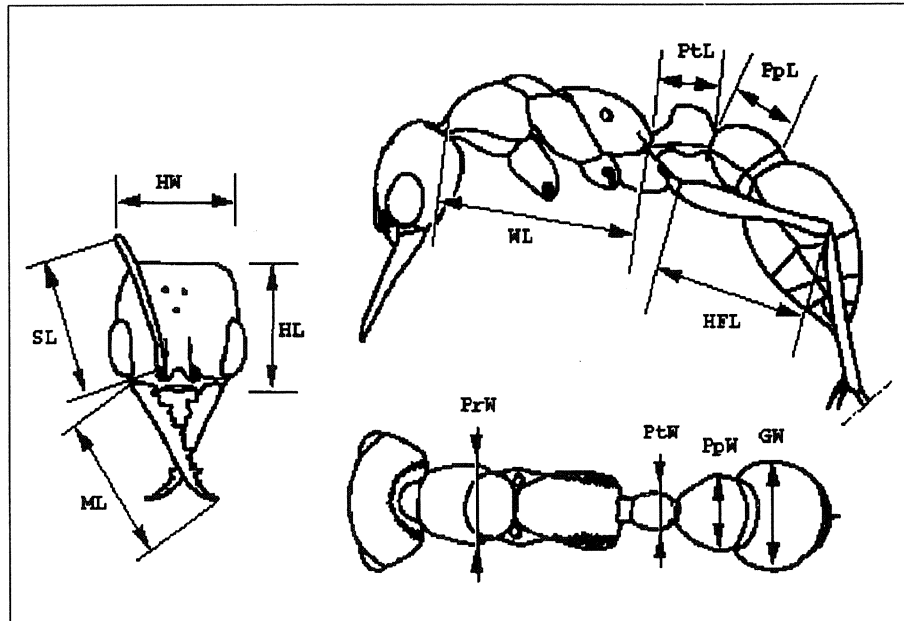


Figure 1: Variables of measurements (*Myrmecia pilosula*).

value in the Table by 20, we can obtain the measurements in the unit of 'mm'. "Species Name with Serial Number" in Table 4 is two digit serial number within species and the species name. Note that the measured specimens were already labelled by previous authors such as Clark (1951), and some of the species names in Table 4 are junior synonyms. To avoid a confusion, a synonymic list is summarized in Table 5 according to the latest taxonomic position. For full taxonomic treatments, see Ogata and Taylor (1991).

Figure 2 illustrate the relationships among selected variables; measurements (HL, SL and PpW; lower left) and indices (MI, PpI and PDI; upper right).

Based on the raw measurement data sets, 10 indices shown in Table 2 may be generated as ratio variables and eliminate size factor. As an example, Table 6 shows the results of classification of 5 groups by a posterior probability of fitted mixture model based on individuals (see Nakamura, 1995a, b), and the classification depended on three indices MI, PpI and PDI except for *urens* species group. The taxonomic information and methods for variable selection are useful for pre-analysis among phylogenetic relationships of the species groups.

Species Group No.	Species Group	Number of Species	Number of Individuals
1	<i>aberrans</i>	6	14
2	<i>cephalotes</i>	3	12
3	<i>gulos</i>	37	123
4	<i>mandibularis</i>	7	17
5	<i>nigrocincta</i>	3	9
6	<i>picta</i>	2	6
7	<i>pilosula</i>	15	42
8	<i>tepperi</i>	5	18
9	<i>urens</i>	4	11
Total		82	252

Table 1: Numbers of species and individuals for each species group.

Address for correspondence: N. Nakamura, 1-1, Jiyugaoka, Yahatanishi, Kitakyushu 807, JAPAN, Phone: +81-93-691-3331, Facsimile: +81-93-692-3245, E-mail: nagatomo@kwuc.ac.jp.

Acknowledgements: The work of measurements was carried out at CSIRO Division of Entomology, Australia, as a part of a project funded by the Australian Biological Resources Study (ABRS) Participatory Program, and this study was partially supported by the Institute of Statistical Mathematics Cooperative Research Program (93-ISM-CRP-A84 and 94-ISM-CRP-A57). We are grateful to Drs. M. J. Whitten (Chief, CSIRO Division of Entomology) and J. F. Lawrence (CSIRO Division of Entomology). The authors thank an anonymous referee for his/her useful suggestion and comments.

References

- [1] Clark, J. (1951), *The Formicidae of Australia, Volume I: subfamily Myrmeciinae*, Melbourne, CSIRO
- [2] Nakamura, N. (1995a), "Statistical Evaluation of Clustering Procedures and Its Applications", Ph. D. Thesis (In Japanese).
- [3] Nakamura, N. (1995b), "Multivariate Normal Mixture Model and the Clustering Procedure", *Bulletin of The Computational Statistics of Japan*, 8, 117-133, (In Japanese).
- [4] Ogata, K. (1991), "Ants of the genus *Myrmecia* Fabricius: a review of the species groups and their phylogenetic relationships (Hymenoptera: Formicidae: Myrmeciinae)", *Systematic Entomology*, 16, 353-381.
- [5] Ogata, K. and Taylor, W. (1991), "Ants of the Genus *Myrmecia* Fabricius: a Preliminary Review and Key to the Named Species (Hymenoptera: Formicidae: Myrmeciinae)", *Journal of Natural History*, 25, 1623-1673.

Index Name (abbreviation)	Treatment	Valid Number of Individuals
Cephalic Index (CI)	HW/HL $\times$ 100	242
Scape Index (SI)	SL/HW $\times$ 100	233
Mandibulo-cephalic Index (MI)	ML/HL $\times$ 100	233
Postpetiolar Index (PpI)	PpW/GW $\times$ 100	244
Hind Femoral Index (HFI)	HFL/PrW $\times$ 100	124
Petiolar Dimension Index (PDI)	PtW/PtL $\times$ 100	235
Postpetiolar Dimension Index (PpDI)	PpW/PpL $\times$ 100	233
Petiolar Index (PI)	PtW/GW $\times$ 100	237
Pronotal Index (PrI)	PrW/GW $\times$ 100	235
Cephalo-gastral Index (CGI)	HW/GW $\times$ 100	240

Table 2: Ratio variables.

	Number of Individuals	Missing	Average	Standard Deviation	Coefficient of Variation	Minimum	Maximum
HW	247	5	59.0	14.7	24.9	20	95
HL	242	10	58.2	15.3	26.2	23	90
SL	233	19	57.9	22.2	38.3	19	103
ML	233	19	59.4	17.7	29.8	21	100
WL	114	138	78.0	12.7	32.2	41	97
PrW	238	14	39.4	8.54	21.7	15	60
HFL	125	127	62.1	14.2	22.9	29	95
PtW	240	12	24.6	4.38	17.8	9	33
PtL	235	17	33.9	12.0	35.4	14	67
PpW	245	7	33.4	6.45	19.9	15	48
PpL	234	18	26.2	5.84	25.2	10	42
GW	245	7	57.8	15.3	26.5	22	93

Table 3: Basic statistics.

Table 4: *Myrmecia* measurement data (1=HW, 2=HL, 3=SL, 4=ML, 5=WL, 6=PrW, 7=HFL, 8=PtW, 9=PtL, 10=PpW, 11=PpL, 12=GW).

Individual Code	Species Name with Serial Number	Species Group Name	Measurements											
			1	2	3	4	5	6	7	8	9	10	11	12
1	01aberrans	<i>aberrans</i>	59	58	40	47	84	40	60	20	25	34	26	50
2	02aberrans	<i>aberrans</i>	53	53	37	40	77	34	58	18	22	30	24	45
3	03aberrans	<i>aberrans</i>	63	61	41	48	89	40	65	22	25	38	28	55
4	01excavata	<i>aberrans</i>	57	55	37	44	82	39	56	20	25	33	22	48
5	02excavata	<i>aberrans</i>	53	52	35	41	75	35	50	17	22	30	21	43
6	01formosa	<i>aberrans</i>	62	60	41	47	87	40	65	20	24	34	25	48
7	02formosa	<i>aberrans</i>	66	65	41	49	91	43	68	23	27	40	27	55
8	01froggatti	<i>aberrans</i>	66	61	45	51	95	42	66	23	27	40	25	55
9	01maura	<i>aberrans</i>	69	67	43	50	90	44	69	25	27	40	26	59
10	02maura	<i>aberrans</i>	62	62	41	45	88	40	65	21	25	37	22	55
11	03maura	<i>aberrans</i>	65	63	40	50	—	—	—	—	—	38	—	56
12	01nobilis	<i>aberrans</i>	69	65	45	49	95	45	67	23	26	40	27	61
13	02nobilis	<i>aberrans</i>	55	55	40	42	89	42	62	22	22	40	26	59
14	03nobilis	<i>aberrans</i>	60	57	40	45	86	39	65	18	24	35	24	50
15	01cephalotes	<i>cephalotes</i>	57	50	46	51	93	40	69	23	28	42	23	59
16	02cephalotes	<i>cephalotes</i>	51	—	46	—	—	—	—	—	—	—	—	—
17	03cephalotes	<i>cephalotes</i>	61	—	50	—	—	—	—	—	—	—	—	—
18	01callima	<i>cephalotes</i>	53	45	46	52	84	36	65	20	24	31	22	46
19	02callima	<i>cephalotes</i>	49	44	45	49	84	33	65	19	24	29	20	45
20	03callima	<i>cephalotes</i>	55	48	47	52	85	36	67	20	25	32	21	49
21	04callima	<i>cephalotes</i>	53	47	45	52	84	37	66	20	25	31	19	48
22	01hilli	<i>cephalotes</i>	55	50	50	54	92	38	60	22	25	35	24	50
23	02hilli	<i>cephalotes</i>	55	50	49	53	91	39	70	20	26	35	25	50
24	03hilli	<i>cephalotes</i>	68	58	55	59	—	49	80	29	32	47	30	67
25	04hilli	<i>cephalotes</i>	63	—	53	—	—	—	—	—	—	—	—	—
26	05hilli	<i>cephalotes</i>	56	—	50	—	—	—	—	—	—	—	—	—
27	01 analis	<i>gulosa</i>	63	62	73	70	—	42	—	22	38	31	27	65
28	02 analis	<i>gulosa</i>	63	65	75	63	—	42	—	20	38	32	25	66
29	03 analis	<i>gulosa</i>	61	60	74	68	—	41	—	20	36	29	26	61
30	04 analis	<i>gulosa</i>	67	65	75	75	—	43	—	22	42	31	30	65
31	05 analis	<i>gulosa</i>	63	64	77	71	—	41	—	22	39	31	27	65
32	06 analis	<i>gulosa</i>	65	66	78	73	—	42	—	23	39	31	30	64
33	01arnoldi	<i>gulosa</i>	75	77	98	85	—	46	—	20	50	31	30	74
34	02arnoldi	<i>gulosa</i>	74	74	90	80	—	47	—	23	50	33	28	75
35	03arnoldi	<i>gulosa</i>	73	74	95	82	—	46	—	22	49	33	27	72
36	04arnoldi	<i>gulosa</i>	71	72	92	80	—	43	—	22	47	31	27	68
37	01atrata	<i>gulosa</i>	70	73	90	80	—	44	—	21	51	32	28	68
38	02atrata	<i>gulosa</i>	69	71	90	76	—	46	—	23	50	34	32	69
39	03atrata	<i>gulosa</i>	62	64	83	72	—	36	—	17	44	27	27	57
40	04atrata	<i>gulosa</i>	74	77	94	85	—	47	—	24	55	35	35	73
41	05atrata	<i>gulosa</i>	51	55	70	61	—	34	—	16	36	26	25	55
42	06atrata	<i>gulosa</i>	83	84	103	99	—	52	—	25	60	40	37	85
43	01auriventris	<i>gulosa</i>	70	66	65	68	—	46	—	22	33	35	26	69
44	02auriventris	<i>gulosa</i>	66	62	60	60	—	41	—	20	32	33	24	62

45	03auriventris	<i>gulosa</i>	70	67	65	66	–	44	–	21	32	33	26	66
46	04auriventris	<i>gulosa</i>	64	60	59	61	–	40	–	18	31	30	24	57
47	05auriventris	<i>gulosa</i>	67	65	65	66	–	44	–	20	36	34	29	64
48	06auriventris	<i>gulosa</i>	75	71	66	70	–	51	–	24	40	38	30	71
49	01basirufa	<i>gulosa</i>	77	80	99	83	–	45	–	27	53	38	32	75
50	01brevinoda	<i>gulosa</i>	65	64	67	67	–	41	–	24	37	36	27	67
51	02brevinoda	<i>gulosa</i>	46	48	49	49	91	30	72	18	29	–	–	–
52	03brevinoda	<i>gulosa</i>	66	66	68	75	–	44	–	25	37	37	28	67
53	01cardigaster	<i>gulosa</i>	54	57	55	62	–	36	–	16	23	26	24	56
54	01comata	<i>gulosa</i>	83	81	76	85	–	56	–	29	48	42	35	80
55	02comata	<i>gulosa</i>	74	76	73	80	–	52	–	27	43	39	31	71
56	01decipians	<i>gulosa</i>	82	76	85	87	–	55	–	28	48	42	33	78
57	02decipians	<i>gulosa</i>	72	72	80	82	–	47	–	25	45	38	28	73
58	03decipians	<i>gulosa</i>	90	86	90	93	–	58	–	30	50	48	37	87
59	01desertorum	<i>gulosa</i>	78	81	100	86	–	52	–	26	53	38	30	78
60	02desertorum	<i>gulosa</i>	70	74	90	79	–	45	–	24	50	34	28	74
61	03desertorum	<i>gulosa</i>	80	82	97	87	–	55	–	29	56	42	33	84
62	04desertorum	<i>gulosa</i>	88	87	–	–	–	54	–	29	65	42	36	85
63	05desertorum	<i>gulosa</i>	87	90	–	–	–	55	–	30	67	40	36	85
64	06desertorum	<i>gulosa</i>	66	70	90	79	–	41	–	21	47	33	30	67
65	07desertorum	<i>gulosa</i>	52	55	71	60	–	35	–	17	40	27	20	55
66	01dimidiata	<i>gulosa</i>	80	83	–	95	–	51	–	26	56	40	33	79
67	02dimidiata	<i>gulosa</i>	75	78	93	87	–	48	–	24	53	36	31	73
68	03dimidiata	<i>gulosa</i>	75	77	91	83	–	45	–	24	54	35	30	74
69	01esuriens	<i>gulosa</i>	58	59	64	57	–	40	88	25	31	39	30	64
70	02esuriens	<i>gulosa</i>	64	66	67	59	–	43	90	28	30	44	33	70
71	03esuriens	<i>gulosa</i>	55	58	60	50	–	38	63	25	30	39	30	61
72	04esuriens	<i>gulosa</i>	69	67	65	60	–	50	92	30	36	47	37	75
73	05esuriens	<i>gulosa</i>	53	53	57	48	–	40	75	23	28	35	27	58
74	01fasciata	<i>gulosa</i>	68	72	91	75	–	44	–	24	49	34	32	69
75	02fasciata	<i>gulosa</i>	72	77	93	84	–	46	–	25	53	35	33	73
76	03fasciata	<i>gulosa</i>	78	81	97	86	–	50	–	27	56	40	34	80
77	01forceps	<i>gulosa</i>	95	90	–	–	–	52	–	26	55	40	35	82
78	02forceps	<i>gulosa</i>	84	81	98	100	–	45	–	24	48	35	33	72
79	03forceps	<i>gulosa</i>	72	68	86	79	–	40	–	20	40	31	25	68
80	01forficata	<i>gulosa</i>	69	68	67	67	–	45	–	28	44	40	30	75
81	02forficata	<i>gulosa</i>	60	62	67	65	–	40	–	22	39	34	25	65
82	01fulgida	<i>gulosa</i>	87	88	–	95	–	57	–	30	57	42	37	87
83	02fulgida	<i>gulosa</i>	79	80	93	80	–	52	–	27	49	38	33	78
84	03fulgida	<i>gulosa</i>	88	88	–	92	–	60	–	30	57	43	38	90
85	04fulgida	<i>gulosa</i>	85	85	–	–	–	–	–	–	–	42	–	89
86	05fulgida	<i>gulosa</i>	82	83	–	–	–	–	–	–	–	39	–	82
87	01gratiosa	<i>gulosa</i>	68	73	94	81	–	44	–	24	51	35	30	70
88	02gratiosa	<i>gulosa</i>	69	73	91	82	–	45	–	24	52	35	29	70
89	01gulosa	<i>gulosa</i>	77	79	88	82	–	52	–	25	47	40	34	73
90	01hirsuta	<i>gulosa</i>	79	78	93	92	–	52	–	28	50	40	31	75
91	02hirsuta	<i>gulosa</i>	62	63	76	74	–	40	–	24	43	34	27	65
92	03hirsuta	<i>gulosa</i>	59	61	75	68	–	40	–	21	39	29	26	57
93	01loginodis	<i>gulosa</i>	77	76	80	85	–	48	–	26	48	39	31	74
94	02loginodis	<i>gulosa</i>	55	55	57	58	–	35	–	20	35	29	23	58
95	03loginodis	<i>gulosa</i>	80	79	83	90	–	50	–	26	48	40	35	75
96	01midas	<i>gulosa</i>	64	60	60	63	–	39	92	21	33	35	23	63

97	02midas	<i>gulos</i>	59	58	56	60	–	38	87	20	30	31	23	57
98	03midas	<i>gulos</i>	58	57	58	55	–	38	–	20	–	–	–	–
99	04midas	<i>gulos</i>	–	–	–	–	–	–	–	20	32	32	26	61
100	01minuscula	<i>gulos</i>	68	69	75	70	–	42	–	21	38	34	30	62
101	01mjobergi	<i>gulos</i>	70	83	99	–	–	43	–	25	55	40	33	79
102	01nigriceps	<i>gulos</i>	75	79	98	90	–	50	–	26	55	36	35	75
103	02nigriceps	<i>gulos</i>	74	80	97	87	–	49	–	25	53	35	33	74
104	01nigriscapa	<i>gulos</i>	69	71	80	76	–	45	–	25	41	36	29	74
105	01pavida	<i>gulos</i>	75	81	–	91	–	50	–	25	55	37	34	72
106	02pavida	<i>gulos</i>	61	64	82	66	–	40	–	20	43	28	26	59
107	01picticeps	<i>gulos</i>	68	66	72	64	–	43	–	24	36	36	25	71
108	02picticeps	<i>gulos</i>	73	71	77	70	–	47	–	25	40	38	27	75
109	03picticeps	<i>gulos</i>	69	68	74	66	–	45	–	24	38	38	25	73
110	01pulchra	<i>gulos</i>	73	68	68	63	–	45	95	25	36	39	32	70
111	02pulchra	<i>gulos</i>	68	65	65	62	–	45	92	23	32	38	30	69
112	03pulchra	<i>gulos</i>	77	73	75	71	–	49	–	28	38	43	37	77
113	04pulchra	<i>gulos</i>	63	60	64	59	–	42	90	22	30	35	30	63
114	05pulchra	<i>gulos</i>	58	58	57	54	–	37	82	20	29	35	29	61
115	06pulchra	<i>gulos</i>	74	70	73	68	–	45	–	25	38	42	35	74
116	01pyriformis	<i>gulos</i>	63	63	69	73	–	40	–	22	41	32	30	65
117	02pyriformis	<i>gulos</i>	67	70	71	75	–	45	–	25	45	35	31	70
118	03pyriformis	<i>gulos</i>	76	75	80	84	–	50	–	29	47	39	34	71
119	04pyriformis	<i>gulos</i>	60	60	67	62	–	39	–	20	38	30	27	58
120	01regularis	<i>gulos</i>	58	59	61	65	–	38	–	20	35	30	25	60
121	02regularis	<i>gulos</i>	59	60	63	63	–	38	–	20	36	30	26	58
122	03regularis	<i>gulos</i>	66	65	68	67	–	42	–	22	40	34	30	65
123	04regularis	<i>gulos</i>	67	66	70	68	–	44	–	23	40	35	28	66
124	01rowlandi	<i>gulos</i>	58	63	60	69	–	40	–	18	36	30	25	60
125	02rowlandi	<i>gulos</i>	65	67	64	75	–	44	–	20	38	34	28	65
126	03rowlandi	<i>gulos</i>	66	68	65	77	–	45	–	20	40	33	29	65
127	01rubripes	<i>gulos</i>	80	80	100	88	–	50	–	25	50	38	33	85
128	02rubripes	<i>gulos</i>	75	74	99	80	–	46	–	20	49	32	34	70
129	03rubripes	<i>gulos</i>	70	69	95	78	–	45	–	20	46	31	31	69
130	04rubripes	<i>gulos</i>	70	71	94	77	–	43	–	20	47	30	30	67
131	01rufinodis	<i>gulos</i>	67	70	85	72	–	44	–	20	45	30	26	68
132	02rufinodis	<i>gulos</i>	41	47	52	41	80	27	80	14	29	20	19	–
133	03rufinodis	<i>gulos</i>	44	48	59	43	–	30	82	13	–	20	–	50
134	04rufinodis	<i>gulos</i>	80	80	96	82	–	52	–	26	51	39	32	84
135	01simillima	<i>gulos</i>	83	81	89	91	–	50	–	25	46	40	34	73
136	02simillima	<i>gulos</i>	85	82	87	90	–	51	–	27	46	40	32	78
137	03simillima	<i>gulos</i>	50	51	55	49	94	32	82	16	27	24	22	51
138	01suttoni	<i>gulos</i>	80	82	100	85	–	51	–	26	57	37	35	85
139	02suttoni	<i>gulos</i>	88	88	–	95	–	56	–	33	60	45	42	93
140	03suttoni	<i>gulos</i>	71	75	94	82	–	45	–	23	51	–	30	70
141	04suttoni	<i>gulos</i>	–	–	–	–	–	–	–	–	–	35	–	80
142	01tarsata	<i>gulos</i>	79	78	80	84	–	55	–	30	46	45	40	83
143	02tarsata	<i>gulos</i>	–	–	–	–	–	51	–	30	45	45	38	80
144	03tarsata	<i>gulos</i>	74	73	77	76	–	53	–	27	–	42	–	77
145	04tarsata	<i>gulos</i>	65	64	73	64	–	46	–	25	40	36	30	67
146	05tarsata	<i>gulos</i>	51	53	57	52	–	34	82	19	32	29	25	52
147	06tarsata	<i>gulos</i>	90	83	–	–	–	–	–	32	50	47	37	85
148	01vindex	<i>gulos</i>	75	76	98	86	–	49	–	26	54	39	31	80

149	02vindex	<i>gulosa</i>	77	78	99	86	–	49	–	24	53	36	31	74
150	01fulviculis	<i>mandibularis</i>	48	48	40	50	78	36	55	22	28	34	28	41
151	02fulviculis	<i>mandibularis</i>	50	48	40	50	80	37	57	24	30	35	28	45
152	01fulvipes	<i>mandibularis</i>	57	49	43	53	88	42	62	25	28	39	30	52
153	02fulvipes	<i>mandibularis</i>	58	50	44	56	89	–	62	25	27	39	30	53
154	01gilberti	<i>mandibularis</i>	54	50	41	50	89	39	62	25	30	38	25	50
155	02gilberti	<i>mandibularis</i>	52	49	40	49	85	38	55	23	27	35	25	48
156	01luteiforceps	<i>mandibularis</i>	49	45	38	50	85	35	55	20	25	34	22	45
157	02luteiforceps	<i>mandibularis</i>	50	46	38	52	82	36	54	22	27	34	22	45
158	01mandibularis	<i>mandibularis</i>	63	56	51	62	97	45	73	28	34	43	33	58
159	02mandibularis	<i>mandibularis</i>	57	50	48	59	92	42	64	23	29	38	30	50
160	03mandibularis	<i>mandibularis</i>	50	45	40	54	78	35	55	18	25	32	25	45
161	04mandibularis	<i>mandibularis</i>	61	54	50	63	94	44	69	25	31	40	30	53
162	01potteri	<i>mandibularis</i>	51	48	40	47	83	39	60	22	25	34	22	47
163	02potteri	<i>mandibularis</i>	51	48	39	46	85	38	59	23	27	36	23	47
164	01pilventris	<i>mandibularis</i>	59	54	49	55	95	41	66	25	30	42	30	55
165	02pilventris	<i>mandibularis</i>	64	58	50	59	–	48	73	28	34	47	34	61
166	03pilventris	<i>mandibularis</i>	42	39	33	40	56	30	45	19	21	28	21	40
167	01flammicollis	<i>nigrocincta</i>	46	49	55	53	94	30	64	15	33	22	20	43
168	02flammicollis	<i>nigrocincta</i>	55	56	65	59	–	35	93	19	37	27	25	50
169	01nigrocincta	<i>nigrocincta</i>	45	48	54	51	93	32	80	14	28	24	22	41
170	02nigrocincta	<i>nigrocincta</i>	54	55	60	57	–	37	92	19	32	31	26	50
171	03nigrocincta	<i>nigrocincta</i>	43	46	51	47	90	28	80	15	27	23	20	38
172	01petiolata	<i>nigrocincta</i>	44	45	–	49	85	30	–	15	26	23	20	40
173	02petiolata	<i>nigrocincta</i>	42	45	50	49	87	30	–	14	27	22	19	40
174	03petiolata	<i>nigrocincta</i>	45	45	50	49	–	30	72	–	–	21	–	40
175	04petiolata	<i>nigrocincta</i>	41	43	47	47	–	29	70	13	23	20	17	37
176	01fucosa	<i>picta</i>	42	45	34	41	69	27	54	18	25	25	18	40
177	02fucosa	<i>picta</i>	45	45	36	43	73	30	56	19	26	27	23	45
178	01picta	<i>picta</i>	44	47	37	43	78	30	59	18	22	28	20	43
179	02picta	<i>picta</i>	43	45	36	40	74	30	57	20	25	29	22	45
180	03picta	<i>picta</i>	45	47	36	42	75	32	50	20	–	30	–	44
181	04picta	<i>picta</i>	38	40	33	38	–	25	51	16	23	25	19	38
182	01apicalis	<i>pilosula</i>	44	44	45	43	83	32	62	15	23	27	17	48
183	01celaena	<i>pilosula</i>	42	39	35	40	64	29	53	16	19	26	18	38
184	01chasei	<i>pilosula</i>	60	51	50	52	96	45	81	27	30	45	28	67
185	02chasei	<i>pilosula</i>	51	46	48	50	85	38	74	20	25	32	20	49
186	01chrysogaster	<i>pilosula</i>	44	42	32	39	75	33	48	18	20	30	20	43
187	01cydista	<i>pilosula</i>	43	41	36	39	70	31	–	19	22	30	18	43
188	01dispar	<i>pilosula</i>	49	46	43	48	72	33	57	19	23	29	19	45
189	02dispar	<i>pilosula</i>	49	48	44	43	75	32	58	20	24	30	20	45
190	03dispar	<i>pilosula</i>	43	43	42	44	71	29	–	18	21	26	18	41
191	01elegans	<i>pilosula</i>	51	51	45	55	87	37	65	23	25	35	24	50
192	02elegans	<i>pilosula</i>	44	43	40	50	74	30	59	20	21	29	20	42
193	03elegans	<i>pilosula</i>	42	43	40	48	72	30	56	19	20	28	20	40
194	01harderi	<i>pilosula</i>	46	44	38	44	75	34	59	20	24	33	22	47
195	02harderi	<i>pilosula</i>	54	50	42	47	84	37	62	24	26	37	23	51
196	03harderi	<i>pilosula</i>	47	45	38	42	73	33	–	19	23	32	22	47
197	01ludlowi	<i>pilosula</i>	64	58	50	55	96	48	80	29	28	47	30	66
198	02ludlowi	<i>pilosula</i>	49	46	45	47	84	37	70	22	23	34	23	50
199	03ludlowi	<i>pilosula</i>	57	51	50	52	92	40	75	25	28	40	26	58
200	01michaelseni	<i>pilosula</i>	52	49	40	49	86	38	61	20	25	39	25	54

201	02michaelseni	<i>pilosula</i>	49	45	38	47	82	37	55	19	25	34	26	47
202	03michaelseni	<i>pilosula</i>	50	45	–	–	–	35	–	19	26	35	24	47
203	04michaelseni	<i>pilosula</i>	55	50	43	50	88	41	63	22	27	40	28	55
204	01occidentalis	<i>pilosula</i>	48	45	44	45	78	32	59	20	25	30	21	45
205	02occidentalis	<i>pilosula</i>	49	45	43	44	80	35	56	20	26	30	23	45
206	03occidentalis	<i>pilosula</i>	46	45	40	43	75	33	57	–	–	30	–	47
207	04occidentalis	<i>pilosula</i>	47	–	45	–	–	–	–	–	–	31	–	48
208	05occidentalis	<i>pilosula</i>	47	46	–	–	–	–	–	19	–	30	–	46
209	01opaca	<i>pilosula</i>	44	42	40	42	72	31	53	18	21	28	20	42
210	01pilosula	<i>pilosula</i>	–	–	–	–	79	35	60	20	25	30	21	46
211	02pilosula	<i>pilosula</i>	37	37	38	40	69	27	50	15	20	24	18	37
212	03pilosula	<i>pilosula</i>	48	44	42	45	80	35	57	20	25	30	24	49
213	04pilosula	<i>pilosula</i>	56	55	48	53	89	40	69	23	27	35	27	55
214	05pilosula	<i>pilosula</i>	58	55	49	55	92	42	70	23	30	36	26	56
215	01rugosa	<i>pilosula</i>	55	50	45	49	90	40	65	25	27	42	28	55
216	02rugosa	<i>pilosula</i>	50	48	40	46	82	38	60	22	25	38	25	50
217	03rugosa	<i>pilosula</i>	43	40	35	40	73	34	50	18	22	31	21	43
218	01varians	<i>pilosula</i>	42	40	38	46	72	30	57	17	23	25	20	39
219	02varians	<i>pilosula</i>	48	47	41	49	78	34	62	21	26	30	25	44
220	03varians	<i>pilosula</i>	43	43	38	45	75	30	57	18	24	27	23	40
221	04varians	<i>pilosula</i>	47	46	40	47	79	31	62	18	25	27	23	43
222	05varians	<i>pilosula</i>	45	45	39	48	80	32	60	19	25	30	24	44
223	06varians	<i>pilosula</i>	47	46	40	48	78	33	61	19	27	30	25	45
224	01clarki	<i>tepperi</i>	44	43	40	49	72	30	51	20	21	29	20	40
225	02clarki	<i>tepperi</i>	49	47	43	52	78	34	55	23	25	34	23	49
226	01dixoni	<i>tepperi</i>	42	40	31	42	65	30	–	19	20	29	18	41
227	02dixoni	<i>tepperi</i>	46	41	34	42	70	33	43	22	23	30	21	44
228	03dixoni	<i>tepperi</i>	40	38	30	40	59	28	40	17	18	26	16	37
229	01swalei	<i>tepperi</i>	38	35	30	40	56	25	36	15	18	23	18	35
230	02swalei	<i>tepperi</i>	48	44	35	45	72	33	47	22	24	31	23	45
231	03swalei	<i>tepperi</i>	40	38	32	40	61	29	40	19	20	28	22	39
232	04swalei	<i>tepperi</i>	45	42	35	45	–	32	–	20	22	31	20	43
233	05swalei	<i>tepperi</i>	37	34	28	37	55	26	36	15	17	24	16	34
234	06swalei	<i>tepperi</i>	51	49	40	49	83	40	58	24	25	38	26	52
235	01tepperi	<i>tepperi</i>	42	40	36	46	65	29	47	15	21	25	20	37
236	02tepperi	<i>tepperi</i>	–	–	–	–	–	35	–	22	25	32	25	47
237	03tepperi	<i>tepperi</i>	50	49	46	50	82	35	60	22	25	35	24	48
238	04tepperi	<i>tepperi</i>	53	50	45	52	82	36	60	23	25	35	25	50
239	01testaceipes	<i>tepperi</i>	43	40	32	43	65	30	42	19	21	29	20	40
240	02testaceipes	<i>tepperi</i>	44	40	34	43	65	30	44	19	21	30	20	42
241	03testaceipes	<i>tepperi</i>	38	35	30	40	58	27	39	15	18	25	18	36
242	01dichospila	<i>urens</i>	34	33	23	26	62	25	39	15	18	24	15	34
243	02dichospila	<i>urens</i>	30	30	24	28	53	20	39	12	16	18	15	27
244	03dichospila	<i>urens</i>	26	27	21	25	48	18	34	10	14	16	13	24
245	01exigua	<i>urens</i>	23	24	19	23	41	15	29	10	14	16	10	22
246	01infima	<i>urens</i>	30	32	25	27	55	20	43	11	16	22	16	30
247	02infima	<i>urens</i>	30	33	25	28	54	20	43	11	17	22	15	29
248	03infima	<i>urens</i>	32	33	25	30	59	23	–	13	15	23	15	34
249	04infima	<i>urens</i>	34	35	26	31	65	24	–	17	20	28	20	38
250	05infima	<i>urens</i>	20	25	19	21	44	17	–	9	15	15	13	25
251	06infima	<i>urens</i>	25	25	19	24	–	–	–	–	–	17	–	25
252	01rubicunda	<i>urens</i>	25	23	19	24	41	15	29	10	14	17	10	25

Species Group Name	Species Names	Species Group Name	Species Names
aberrans	aberrans	gulosa	rufinodis
	<i>excavata</i> → froggatti		simillima
cephalotes	formosa	mandibularis	<i>suttoni</i> → fulgida
	froggatti		tarsata
	maura		vindex
	nobilis		fulviculis
	callima		fulvipes
	cephalotes		gilberti
	hilli		luteiforceps
			mandibularis
	analisi		piliventris
	arnoldi		potteri
	<i>atrata</i> → pavida		flammicollis
	auriventris		nigrocincta
	<i>basirufa</i> → vindex		petiolata
	brevinoda		fucosa
	<i>cardigaster</i> → rowlandi		picta
	comata		apicalis
	<i>decipians</i> → brevinoda		<i>celaena</i> → harderi
	desertorum		chasei
	dimidiata		chrysogaster
	esuriens		cydista
	<i>fasciata</i> → nigriceps		dispar
	forceps		elegans
	forficata		harderi
	fulgida		ludlowi
	gulosa		michaelseni
	gratiosa		occidentalis
	hirsuta		<i>opaca</i> → occidentalis
	<i>longinodis</i> → brevinoda		pilosula
	midas		rugosa
	minuscule		varians
	mjobergi		clarki
	nigriceps		<i>dixonii</i> → testaceipes
	nigriscapa		swalei
	pavida		tepperi
	pincticeps		testaceipes
	pulchra		dichospila
	pyriformis		exigua
	regularis		infima
	rowlandi		rubicunda
	rubripes		

Table 5: Index to species names in Table 4 (junior synonymic names are shown by italics).

Group	Number of Individuals	Individual Number
1	13	1,2,3,4,5,6,7,8,9,10,12,13,14
2	59	15,18,19,20,21,23,69,70,71,72,73,150,151,152,154 155,156,157,158,159,160,161,162,163,164,165,183 185,186,187,188,189,190,194,195,196,200,201,203 204,205,209,211,212,213,214,215,216,217,218,219 220,221,222,223,229,235,237,241
3	24	22,24,153,166,184,191,192,193,197,198,199,224 225,226,227,228,230,231,232,233,234,238,239,240
4	40	28,33,34,35,36,37,38,39,40,41,42,49,59,60,61,64 65,66,67,68,74,75,76,84,87,88,102,103,105,106,124 128,129,130,131,134,138,148,149,167
5	71	27,29,30,31,32,43,44,45,46,47,48,50,52,53,54,55,56 57,58,78,79,80,81,82,83,89,90,91,92,93,94,95,96,97 100,104,107,108,109,110,111,112,113,114,115,116,117 118,119,120,121,122,123,125,126,127,135,136,137,139 142,145,146,168,169,170,171,172,173,175,182
Unclassified	34	11,16,17,25,26,51,62,63,77,85,86,98,99,101,132,133 140,141,143,144,147,174,176,177,178,179,180,181,202 206,207,208,210,236,242,243,245,246,247,248,249 250,251,252

Table 6: The results of classification by a mixture model.

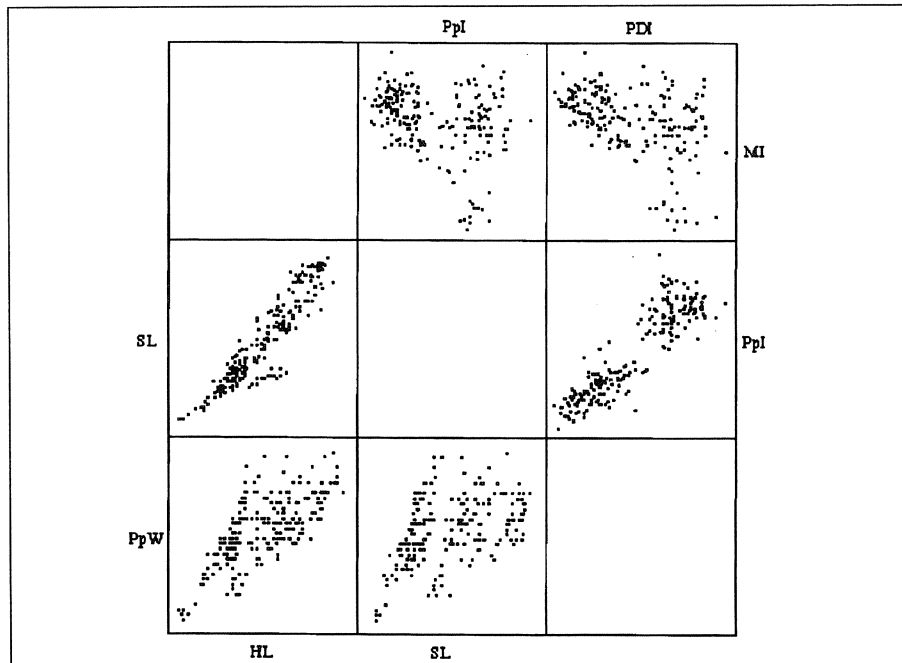


Figure 2: Scatter diagrams of selected variables and indices.

Relative Behavior of a Cohort of Triticale Cultivars Within the Same Equipotential Zone

João Tiago Mexia

Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Portugal

Luís Gusmão and José Baeta

Dept. Genética, Estação Agronómica Nacional, Oeiras, Portugal

Abstract: The data set was extracted from a national network of triticale adaptation trials, in Portugal, and was used to show how a trimming procedure can be applied to the Joint Regression Analysis to forecast the relative grain yield behavior of a cohort of genotypes, under different environments, within the same equipotential zone for relative yield pattern evaluation. The method allows to identify two levels of genotype $\times$ environment interactions: The first level of interaction is expressed in the relative pattern of the joint regressions, after trimming, and can be used to estimate the relative yield behavior, of genotypes, for different environmental indices; the second level is expressed in the non adjustment to the joint regressions of the results of a particular genotype in a particular trial and may reveal a specific interaction, in connection with thorough observations during trial performance.

Key words: Genotype $\times$ environment interaction, grain yield evaluation, joint regression analysis, trimmed regression analysis.

1 The scientific context

The data, presented in this study, were taken out from a network of regional triticale yield trials. The experimental design takes into account the “genotype $\times$ place” and “genotype $\times$ year” interactions; however, as trials were included within the same Equipotential Zone for Cultivar Yield Pat-

tern Evaluation (see Gusmão, Mexia and Gomes, 1989), it is adequate to go on with the Joint Regression Analysis (see Yates and Cochran, 1938), as concluded by Gusmão, Mexia and Gomes (1989). The assumption for this method of analysis is that “place” and “year” can be integrated in one sole variable, the “environmental index” (see Eberhart and Russell, 1966); however, specific genotype $\times$ environment interactions, largely recognized for most of the genotypes, remain to be explained.

Present data were used to show how specific interactions can be identified and, further, used to trim the joint regressions (see Gusmão, Mexia and Baeta, 1992). The relevance of the trimming procedure results both from the detection of particular genotype susceptibility or tolerance/resistance to environmental stresses (specific interactions) and from a better prediction of relative yield behavior at different environmental indices.

The method is based on the analysis of the residues to the first adjusted joint regressions. For each genotype, in each trial, there is a sample of as many residues as the number of replicates (complete blocks). An abnormal behavior of a genotype, in a particular trial, is expressed in the mean rank, of the respective residues, within the family of all residues for that trial. On the other hand, the residues for the several regressions are complementary since the residues for the same block are correspondent. We may so use, for instance, the Friedman test (see Friedman, 1937) to compare the samples of residues.

As a further step on the trimming method, it is convenient to proceed with the fine analysis of the residues. In fact, yield is the result of the interaction of genetic and environmental factors, along the productive life cycle of the plant and, so, the multiplicity of situations that may arise can be thought as a “physical support” to a “Central Limit Theorem”. Thus, it is of paramount relevance to discuss at what level the residues distribution approaches the normality. This study should be undertaken after elimination of the correlation between residues that were induced by the trimming procedure.

2 The data

In Tables 1 and 2 we present the grain yield results for the same cohort of genotypes, respectively for 1990 and 1991.

The present data sub-set was taken out from a network of regional triticale yield trials, carried on by the Portuguese Plant Breeding Station (*Estação Nacional de Melhoramento de Plantas* - ENMP). Each trial was designed in Randomized Complete Blocks, with four replications and a plot area of 10 m² (2 x 5 m).

Soil and crop management were performed according local and seasonal conditions.

Address for correspondence: João Tiago Mexia, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2825 Monte da Caparica, Portugal. Phone: 351 1 2954464, Fax: 351 1 2955641.

References

- [1] Eberhart, S.A. and Russell, W.A. (1966), "Stability Parameters for Comparing Varieties", *Crop Sci.*, 6, 36-40.
- [2] Friedman, M. (1937), "The use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance", *J. Amer. Statist. Assoc.*, 32, 675-701.
- [3] Gusmão, L., Mexia, J.T. and Baeta, J. (1992), "Trimmed Joint Regression: A new Approach to the Joint Regression Analysis for Cultivar Relative-Performance Evaluation, *Theor. Appl. Genet.*, 84, 735-738.
- [4] Gusmão, L., Mexia, J.T. and Gomes, M.L. (1989), "Mapping of Equipotential Zones for Cultivar Yield Pattern Evaluation, *Plant Breeding*, 103, 293-298.
- [5] Yates, F. and Cochran, W.G. (1938), "The Analysis of Groups of Experiments", *J. Agric. Sci.*, 28, 556-580.

Place	Rep.	Genotype (yield - kg ha ⁻¹)						
		Arabian	Borba	Arruda	Crato	TTE8803	TTE8901	TTE8902
Portalegre	I	1533	1833	1600	1600	1867	1933	1333
	II	1500	1533	2067	1667	1800	2067	1800
	III	1333	1533	1733	1867	1867	2267	1667
	IV	1033	1400	1467	1400	1267	1400	1133
Abrantes	I	3185	2988	2720	2637	2226	2851	2155
	II	2438	3113	2952	2619	2304	2702	1637
	III	2405	2827	2625	2720	2351	2351	1720
	IV	2702	2208	2601	3066	2577	3363	2381
Santarém	I	2738	4345	5298	4762	4167	5238	5000
	II	3095	4345	6072	3452	3512	4822	4643
	III	2381	3572	3929	4345	4286	3929	4048
	IV	3274	3452	4643	4048	4464	4822	4524
Chamusca	I	405	935	708	786	952	1101	708
	II	738	976	940	1095	1304	1095	815
	III	1155	1131	702	958	1149	1179	857
	IV	619	250	1226	869	940	869	536
Pegões	I	1810	1643	1833	1685	1595	905	1000
	II	1488	1500	1679	2036	2095	1976	1857
	III	1857	1619	2440	1869	1655	1917	1595
	IV	2083	2119	1738	1917	2155	2274	2012
Benavila	I	1867	1767	1533	1467	1700	2167	1733
	II	1500	1333	1400	1333	1667	2000	1567
	III	1200	1600	1300	1633	1133	1867	1367
	IV	1267	1833	1667	2000	2267	2133	1733
Crato	I	333	667	667	867	733	600	567
	II	667	667	933	1000	833	1033	1033
	III	400	800	600	600	667	900	700
	IV	533	733	700	633	367	500	667
Comenda	I	4881	6012	5929	5476	5953	6774	6548
	II	4881	6845	6667	6381	6369	6607	6988
	III	4941	5655	6548	6072	6250	7345	6310
	IV	4702	6310	6548	6310	6643	6548	5774
Elvas	I	3428	3923	5372	4607	4600	4805	4965
	II	3921	3944	4701	4176	4059	5188	4652
	III	4043	3896	4916	4383	3968	4784	4824
	IV	3781	3623	4135	3925	4175	5374	4481
Revilheira	I	267	400	600	733	667	867	867
	II	67	133	533	467	533	467	267
	III	67	133	267	333	533	867	400
	IV	133	400	267	400	333	400	267
Évora	I	464	667	2012	1875	1464	2024	1446
	II	917	929	1583	1792	2298	2238	2083
	III	1107	1191	2226	2274	2357	2470	2429
	IV	970	1476	3292	3083	2488	2655	2530
Beja	I	3691	2976	5119	4167	6905	5357	4881
	II	2738	3691	4048	3572	4405	4643	4762
	III	3333	4167	4524	4286	3929	4167	4405
	IV	2738	3333	4881	4881	4643	6310	5357
Tavira	I	60	1905	1399	1399	1161	1250	804
	II	119	1369	1905	1756	1399	1369	1042
	III	179	1905	2113	1726	1637	1667	1250
	IV	179	1012	1696	1577	1667	1935	1488

Table 1: Yields results of a network of triticale adaptation trials, for the same cohorte of genotypes, within the same Equipotential Zone (South “Montejunto-Estrela”, Portugal), in the year 1990.

Place	Rep.	Genotype (yield - kg ha ⁻¹)						
		Arabian	Borba	Arruda	Crato	TTE8803	TTE8901	TTE8902
Mirandela	I	3738	2506	3708	3613	3089	4429	2905
	II	4952	4196	4917	4310	3833	3536	3125
	III	3423	2345	2375	3048	3875	4310	2440
	IV	1893	1458	1137	1827	1137	1327	1190
Abrantes	I	3571	3512	2976	3452	3571	3810	3036
	II	3810	2619	3571	2560	2619	4167	2738
	III	2857	2024	2381	2381	1905	1667	2500
	IV	2143	1667	2976	2798	1905	2262	1905
Santarém	I	1190	1190	1190	1786	1190	1548	1190
	II	714	2738	1190	2024	1190	1548	1310
	III	1071	1786	1310	1548	1190	2143	1964
	IV	952	2738	2024	1190	1429	1786	1667
Pegões	I	2440	2143	2500	2500	2262	1964	2143
	II	1667	2440	3214	3095	1905	2440	2024
	III	1905	2321	2679	2440	2619	2738	2024
	IV	2738	3036	2500	2024	2024	2083	1488
Benavila	I	4053	3613	3253	3267	3267	3160	4467
	II	4027	3667	3693	3133	3133	3600	3787
	III	3533	4800	3640	3507	3587	4347	3573
	IV	4093	3587	3547	3240	3333	4173	3293
Comenda	I	6607	5714	5952	6548	6071	5357	4405
	II	6845	6726	7202	7262	6310	6429	5655
	III	5357	4821	5655	6012	5714	5476	5179
	IV	5000	4524	4643	4762	4286	4643	3393
Elvas	I	5295	7108	5948	4795	5513	6537	6928
	II	8787	9119	8580	8111	5283	8471	8888
	III	8452	7448	7256	8167	6919	7308	5681
	IV	6256	6441	7103	7076	6572	9251	7757
Revilheira	I	5267	3520	4215	4163	4080	4661	4595
	II	5124	4285	4479	4733	4400	4185	3744
	III	4053	3273	4163	4148	4129	4369	3577
	IV	5717	4675	4223	4855	4640	5043	4247
Beja	I	4286	3571	3571	3333	3571	3929	3333
	II	4524	3810	3571	5000	4524	4762	3929
	III	5357	4881	4048	4524	3929	4524	4762
	IV	5357	5357	5238	3690	5000	6071	4762
Tavira	I	440	357	2381	2387	1946	4446	3244
	II	952	1690	3036	2452	1988	1720	1940
	III	1012	1369	2804	2399	1667	4964	2470
	IV	476	881	3946	3304	2667	2732	2946

Table 2: Yields results of a network of triticale adaptation trials, for the same cohorte of genotypes, within the same Equipotential Zone (South “Montejunto-Estrela” and “Low Land of the Northeast”, Portugal), in the year 1991.