Fresenius J Anal Chem (1995) 352:431–436

Fresenius' Journal of Analytical Chemistry

Multivariate 3-way data analysis of amino acid patterns of lakes

G. Henrion[1], D. Nass[2], G. Michael[1], R. Henrion[3]

[1]Humboldt University, Institute of Chemistry, Hessische Strasse 1–2, D-10115 Berlin, Germany
[2]Schering AG, Department of Analytical Development, D-13342 Berlin, Germany
[3]Humboldt University, Institute of Applied Mathematics, D-10099 Berlin, Germany

Received: 31 January 1995 / Revised: 3 April 1995 / Accepted: 14 April 1995

Abstract. Time dependent patterns of amino acid concentrations have been studied by HPLC for different lakes of the Berlin area. Data analysis has been performed by conventional principal component analysis as well as by its more recent N-way extension. It turns out that lakes mainly differ by their general amino acid production as a function of time and season. Apart from this, in a single case there occurs a specific pattern which might be related to an exterior influence. This pattern, although clearly detected, has been not stable over time. Measurements are reproducible with respect to time (comparison of two succeeding years) and to position (comparison of isolated parts of a lake).

Introduction

In the late sixties first papers have been published with the goal to analyze concentrations of various amino acids in different lakes or rivers in order to get an indirect measure for their biological balance and state [1–4]. This seemed to be reasonable, because amino acids play a key role in synthesis and decomposition of organic matter and for life in natural systems. It could be expected, that the essential amino acids form different concentration patterns for different organisms (e.g. invertebrates [5], bacteria [6], plankton [7], algae [8]). Besides, they might serve as an indicator for the type and the 'health' of natural waters. Deviations from typical patterns would be referable to stress by human and industrial influence. For the determination of amino acids in natural waters high performance analytical procedures are necessary and have been described in the literature [9–12]. Among them, HPLC holds a dominant position [13–22]. A great problem is, that amino acid concentrations change quickly in time, depending on the competition of synthesis and decomposition within the multiplicity of biological processes, which are subjected, for instance, to changing light and temperature. Consequently, for the purpose of comparability and representativity, samples from different lakes should be taken simultaneously, be mixed from different sampling sites of any lake and be soon sterilized and freezed [23–25]. This makes evident, that characterization of rivers is yet more difficult. But even in the easier case of lakes the resulting data structure is rather complex. The extraction of interpretable hypotheses from multidimensional observations, which are interfered with noise due to the described complications, requires tools from multivariate data analysis [26].

Dedicated to Professor Dr. K. Doerffel at his 70th birthday with respect to his fundamental contributions to chemometrics

Correspondence to: G. Henrion

Experimental

The investigated lakes are typical for the nearer and further surrounding of Berlin: Müggelsee, Dämeritzsee, Langer See, Flakensee, Werlsee, Peetzsee, Möllensee (Fig. 1 a) and Großer Vätersee, Döllnsee, Stechlinsee, Große Fuchskuhle (Fig. 1 b). Große Fuchskuhle is a small clean lake which, for experimental purposes, was divided by foils into four equally sized and isolated sectors without water exchange. The sectors were given different initial conditions.

Water samples from the lakes to be investigated were provided by the 'Institut für Gewässerökologie and Binnenfischerei' (Berlin and Neuglobsow), where other chemical and biological parameters are monitored over long periods. Sampling was carried out under defined conditions and at fixed times. Depending on area, depth and velocity of flow of the respective lakes, parallel samples from different sites were mixed to give a representative average. These were immediately separated from suspended organic and inorganic matter by filtration (0.2 µm cellulose acetate) and sterilized by the addition of mercury chloride solution to stop further processes, in particular the decomposition of amino acids by microorganisms. 100-ml-samples of water were evaporated by freezing and vacuum. The residue was dissolved in 1 ml of buffer solution (19.6 g sodium citrate and 50 ml mercapto-ethanol in 800 ml H_2O, pH adjusted to 2.2 by HCl, addition of 30 ml mercapto-ethanol and filled up to 1000 ml). For derivatization, 10 µl of this solution were mixed with 10 µl of OPA-reagent (60 mg o-phthalic acid aldehyde, 5 ml methanol, 45 ml borate buffer solution, 1 ml mercapto-ethanol adjusted to pH = 10 with NaOH; the borate buffer contained 30.9 g H_3BO_3 in 500 ml H_2O, pH = 9.5 adjusted by NaOH). Separation was achieved on a 250 × 4 mm Hypersil ODS (Octadecylsilane), 5 µm column at 50° C with gradient elution. The com-

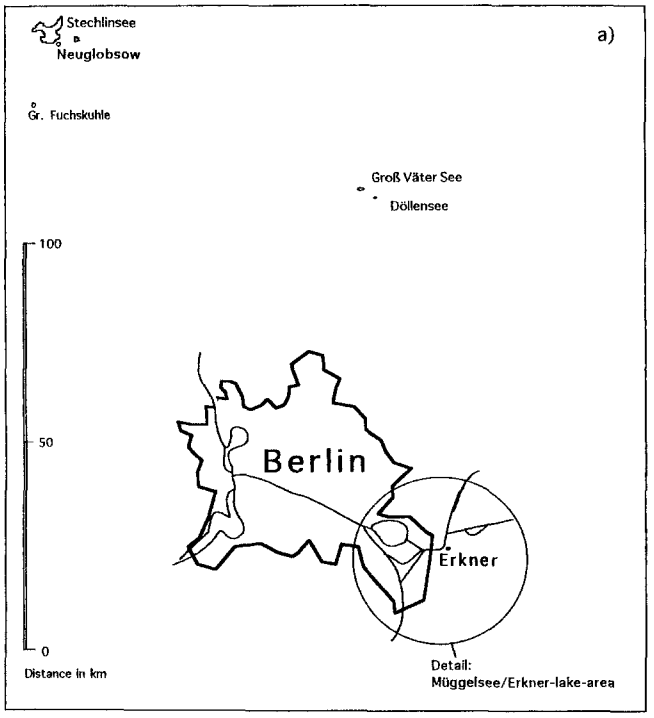

Fig. 1 a, b. Map of **a** the region of investigated lakes and **b** the Berlin area enlarged

position varied from 95% phosphate buffer (2 g $NaH_2PO_4 \cdot 2\ H_2O$, 2250 ml H_2O, 250 ml methanol, adjusted to pH = 8 with NaOH) and 5% eluent solution B (97% methanol and 3% tetrahydrofuran) to 10% phosphate buffer and 90% eluent solution B [27, 28]. The order of retention time of the 15 amino acids separated is listed together with the abbreviations used: asparagic acid (asp), glutamic acid (glu), serine (ser), histidine (his), glycine (gly), threonine (thr), arginine (arg), alanine (ala), tyrosine (tyr), methionine (met), valine (val), phenylalanine (phe), isoleucine (ile), leucine (leu), lysine (lys). The fluorimetric detection was carried out at 330 nm (excitation) and at 455 nm (emission). Figure 2 shows typical chromatograms of samples taken at different seasons from lake Müggelsee as well as a standard solution of the 15 amino acids used for calibration (e.g. 5 nmol/ml).

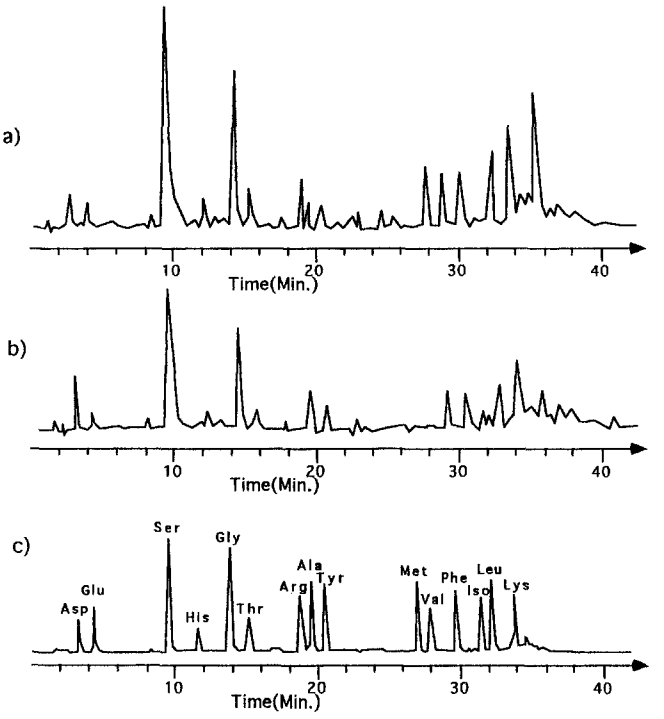

Fig. 2 a–c. Chromatograms of samples from the lake Müggelsee taken **a** in April and **b** in July as well as of a standard solution used for calibration

Results and discussion

The investigations result typically in data tables for each lake where rows correspond to the sampling times and the columns represent the amino acids measured. Table 1 shows a very small part of such measured data for the purpose of illustration. The values recorded are always averages of three parallel estimations. It becomes evident even from this small table, that the total data variation splits into temporal changes and differences in magnitudes between amino acids. From this it follows, that the differences in magnitudes should be removed by appropriate scaling (see below) in order to get time-dependent patterns which are equally influenced by all amino acids. As a first step of evaluation one could trace the time dependent concentrations of each single amino acid, as it is depicted in Fig. 3 for the lake Müggelsee (based on 73 sampling times from October 92 to December 94). Not surprisingly, some periodical behavior can be observed within the whole family of curves (see e.g. the peaks at current weeks 30 and 80 in a one year distance). On the other hand, not all of the amino acids show a similar depen-

Table 1. Small part of a typcial data table obtained. Concentration in ng/l

Sampling date	asp	ser	arg	met
7/12/92	3546	16473	1049	304
21/12/92	9086	15701	563	170
18/1/93	7025	8369	384	37

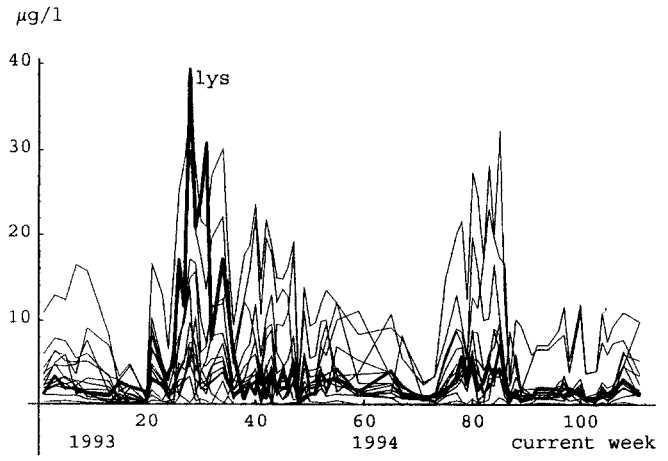

Fig. 3. Concentrations of 15 amino acids in the lake Müggelsee as a function of time

dence on time, as one can see from the lys-curve in Fig. 3. Furthermore, each of the curves is quite noisy due to unavoidable experimental errors. This, altogether, makes it hard to recognize the essential data structures.

Principal component analysis (PCA) is a well known tool of multivariate data analysis which allows an efficient noise reduction and a quick extraction of some main relations contained in the data table. Details of the method can be obtained from standard textbooks as [29, 30]. Although many other methods could be additionally applied, the data evaluation will be restricted to conventional PCA and to its N-way extension (see below). Roughly speaking, PCA explains a major part of correlating data variation by a minor part of independent principal components (PCs). In a PC1 vs. PC2 diagram, these components allow an effective representation of the rows (here: sampling times) and columns (here: amino acids) of a data table. An example is given in Fig. 4 for the lake Stechlinsee. As usual in multivariate data analysis, appropriate scaling is done beforehand. This results in each column of the data table getting equal mean zero and equal variance one. Such scaling prevents some amino acids with high concentrations and great dispersion over the sampling period to dominate the data structure (compare the magnitudes of concentration in Table 1). As a consequence, the samples will be centered around the origin (which is interpretable as an average sample) in all PC-plots. Sometimes it is convenient to discuss PC-plots according to each axis. Accordingly, almost all amino acids (except arg) get positive weights on the PC1-axis in Fig. 4 (top). Hence, one can speak of a general factor of amino acid production. Discussing PC1-coordinates of the consecutively numbered sampling times – which are not equidistant – one recognizes that early times (1–7) result in values in the general factor which are below the average while a sharp maximum occurs at sampling time 11 (recall data scaling). For better visualization the PC1-coordinates can be plotted versus the real sampling time (current week) as in Fig. 4 (bottom). The resulting curve may be understood as an image of the seasonal development of the general amino acid production in the lake Stechlinsee. The vertical axis PC2 in Fig. 4 (top) is clearly dominated by high weights of arg and, to a smaller degree of tyr. This means that, apart from the general factor of the amino acid production, there exists an independent specific factor which is mainly determined by an arg peak at sampling times 4 and 9. Similar to PC1 one could plot the PC2 values over the sampling times.

The PC-plot of Fig. 4 is somehow typical for all the lakes investigated: the main data variance, described by PC1, is due to a temporal dependence of concentrations which is common to amino acids as a whole, but will differ, of course, between the lakes. Besides, there frequently occurs a 'pattern', i.e. an independent function of time for some subgroups. Whether such patterns are typical and reproducible for a fixed lake can only be confirmed by long-term observations. In the present study the maximum range of observations has been recorded for the lake Müggelsee over a period of two years (Fig. 3). The resulting PC-plot in Fig. 5 comprises amino acids only, since the set of 73 sampling times would yield an noninterpretable cloud of points. Instead, the time dependence corresponding to both PCs was plotted directly in Fig. 6. Again, all amino acids have a positive weight on the PC1 axis although with some stronger differentiation compared to Fig. 4. The time dependence of this general factor

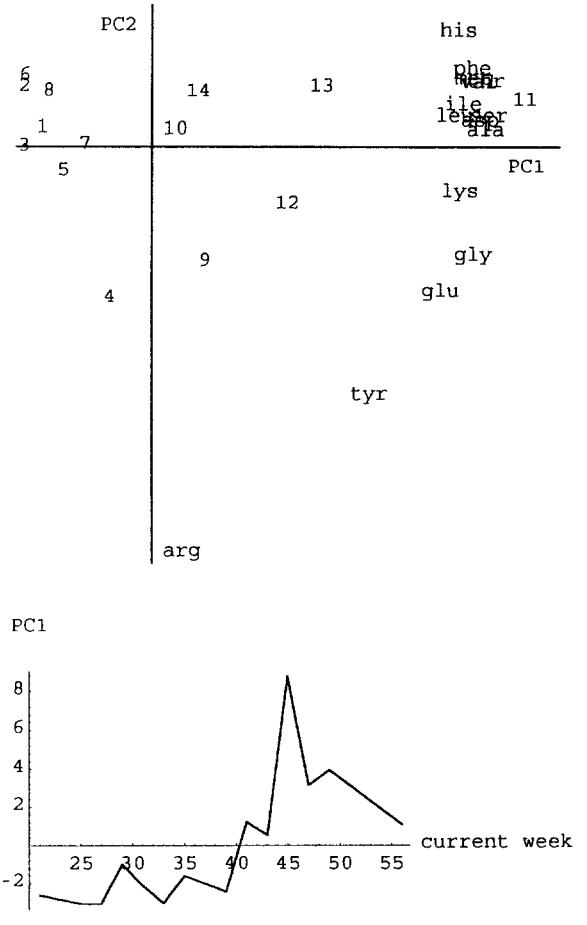

Fig. 4. PC-plot of amino acids and sampling times for the lake Stechlinsee (*top*) and representation of PC1 coordinates of time points versus the sampling times themselves (*bottom*)

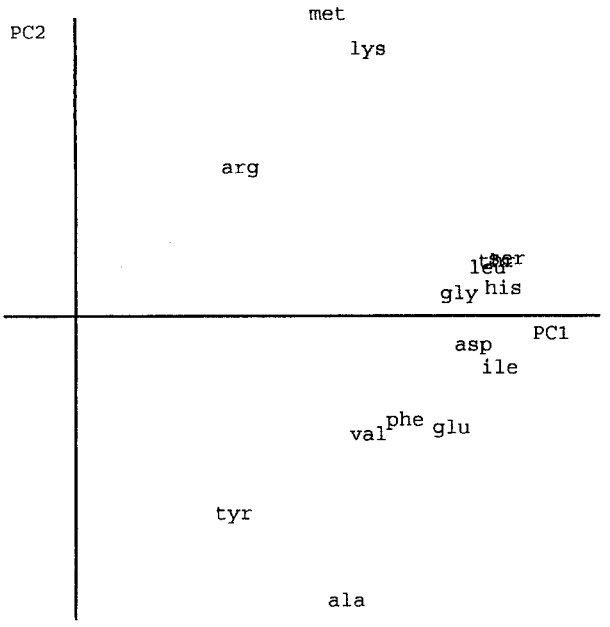

Fig. 5. PC-plot of amino acids for the lake Müggelsee

of amino acid production is shown together with measurement points in Fig. 6 (top). The position of the horizontal axis reflects the average value over the considered period. Not surprisingly, there are clear differences between the lakes Stechlinsee and Müggelsee (Fig. 4 (bottom)). While the first one shows a single pronounced maximum in the late summer (current week 45), several maxima with distinct intensities are observed for the latter. It is well known from monitoring of microorganisms and other parameters [31, 32], that the Müggelsee has a maximum activity in early spring and, to a lower degree, in late summer. This is in loose coincidence with the given curve (current week 25 = April 93, current week 40 = August 93). However, it can be recognized form the picture, that the amino acid production shows some more oscillating behavior with a roughly estimated period of about five weeks. Not only a major part of peaks in the curves is supported by a sufficient number of measurement points to be distinguished from noise, but also the peaks themselves are reproducible one year later. To make this more evident, in Fig. 6 (center) the right part (1994) of the curve above is shifted towards the left by 52 weeks (dashed line). In relation to the measurement error to be expected here, one gets quite a good reproduction, particularly of the monotone decrease from November to March (current weeks 1–20) and of the three peaks occurring in spring time (current weeks 20–35). Concerning the summer and autumn period (current weeks 35–60) the intensities differ significantly but still the correlation is remarkable.

Looking back to Fig. 5, there is a pattern related to the vertical axis PC2. It consists of met, lys, (arg) on the one hand and, negatively correlated, tyr, ala on the other. Time dependence of this pattern, which is independent of the general amino acid production, is illustrated in Fig. 6 (bottom). A clear maximum can be detected at the spring time of the first year (current week 30) while rather a minimum is detected one year later (current week 80–95). While

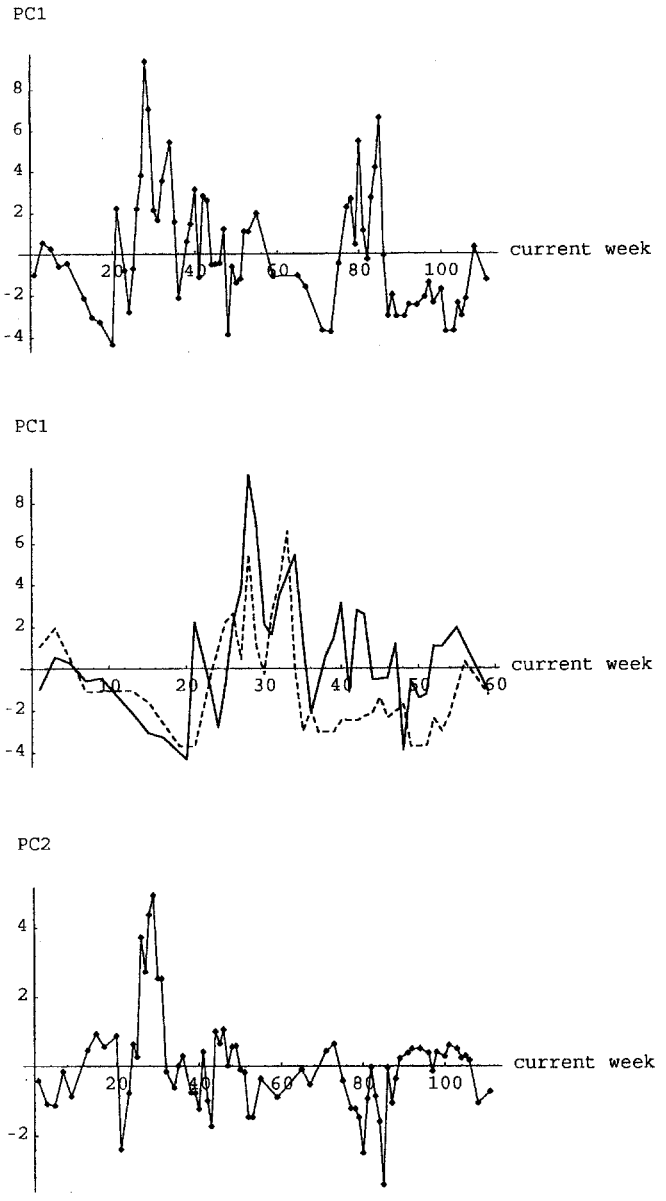

Fig. 6. PC1 coordinates of time points for the lake Müggelsee versus sampling time (*top*), same plot with the right part (second year) shifted into the left part (first year) (*center*) and corresponding plot of PC2 coordinates (*bottom*)

the indicated pattern is not reproducible after one year – and therefore cannot be interpreted as a typical pattern of the lake Müggelsee – its occurrence in the first year is supported by more than 10 measurement points, which precludes it from being some measurement artefact. Perhaps it indicates some exterior influence, but an explicit explanation is not yet known.

Up to now PCA was performed by considering fixed lakes over differing sampling times. Analogously one could fix the sampling time and look for differences of amino acid patterns between the lakes. Then, the lakes rather than the sampling times become the rows of the data tables and a PC-plot should yield, apart from amino acids, a point set corresponding to the lakes. Unfortunately, measurements of a major fraction of lakes was available

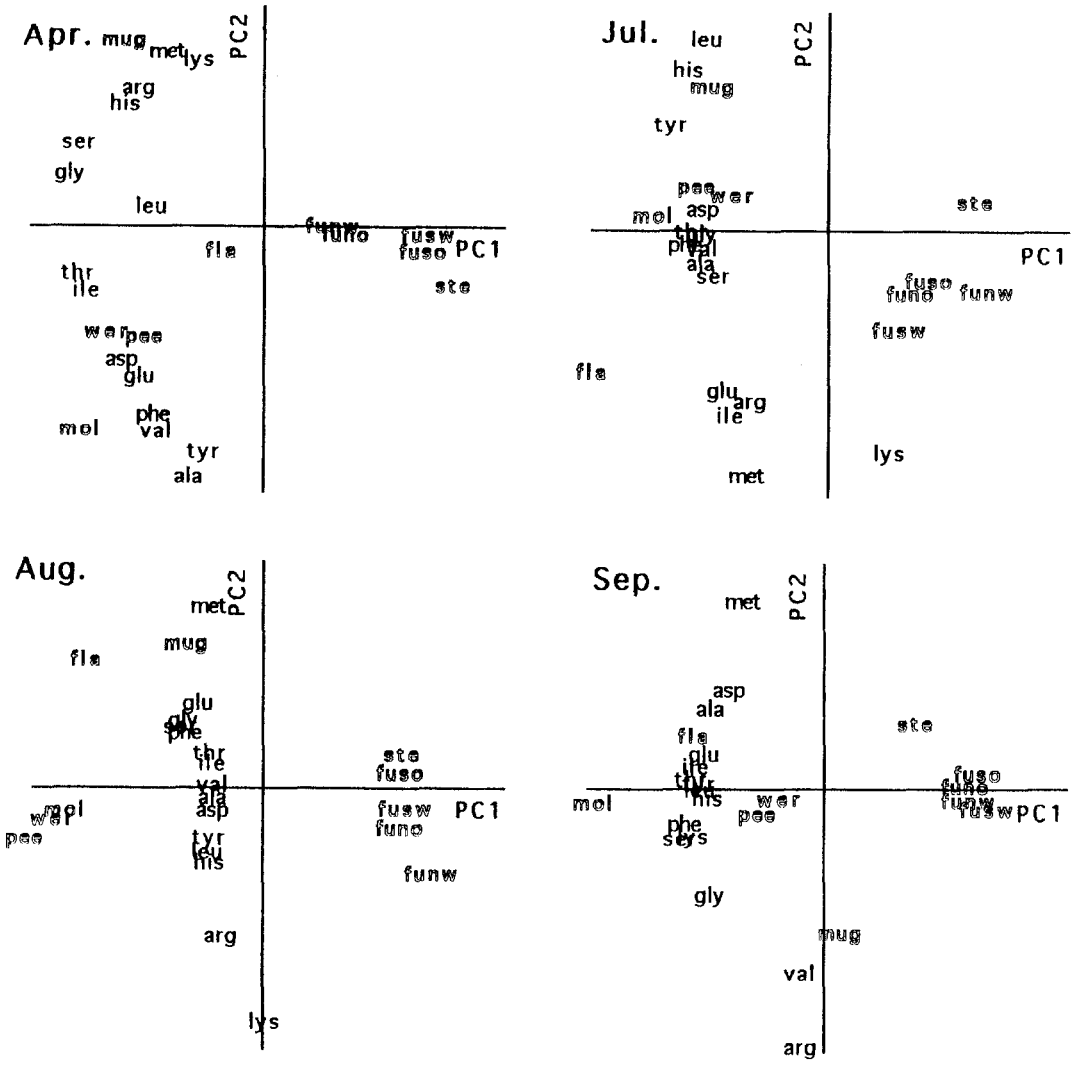

Fig. 7. Separate PC-plots of lakes and amino acids for different sampling times

only for a restricted set of sampling times. Figure 7 contains the PC-plot relating to the data from April, July, August and September. The following fraction of investigated lakes was considered (with abbreviations in parantheses): Flakensee (fla), Werlsee (wer), Peetzsee (pee), Möllensee (mol), Stechlinsee (ste), Müggelsee (mug) and the four parts of Große Fuchskuhle – northeast (fune), northwest (funw), southeast (fuse), southwest (fusw). In any case, almost all amino acids result in equal signs on the first axis confirming PC1 to be a factor of the general amino acid production. Accordingly, the PC1 coordinates of the lakes give a distinction with respect to this factor. As a matter of fact, Stechlinsee and Große Fuchskuhle differ from the other lakes by lower concentrations in all amino acids. On the other hand, the arrangement of lakes and amino acids along the PC2 axis is rather unstable making it complicated to detect interpretable patterns, Anyhow, the vertical position of the Müggelsee is distinguished from the others by differing concentrations in met, lys and some others.

A discussion of separate PC-plots as in Fig. 7 is quite difficult since it is hard to detect factors which are stable over time. The reason is, that the very data arrangement of the present study is not a table, as for a fixed lake or for a

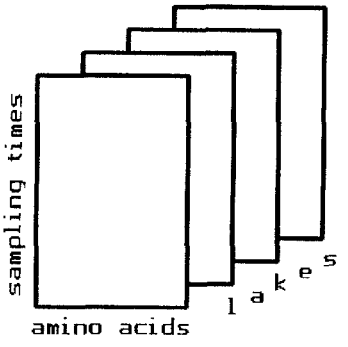

Fig. 8. Sketch of the three-dimensional data structure in the present study

fixed sampling time, but rather something like a data cube, as depicted in Fig. 8. Instead of a single table, one deals with several tables which occur as different slices in a threedimensional array. A simple method to get a joint analysis of all these slices would be to piece them together along one of several possible directions. Then, the application of conventional PCA to this large data table is directly possible. However, while such procedure is sufficient in special cases as for data from image analysis [33], it makes PC diagrams generally hard to interpret due to generating so-called combination modes, the elements of

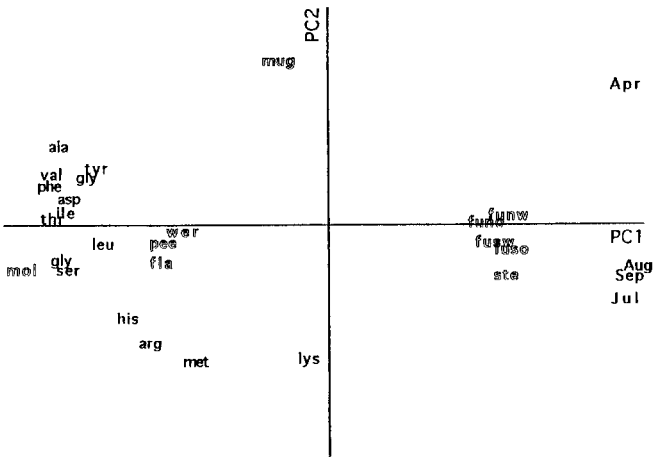

Fig. 9. N-way PC-plot of lakes, amino acids and sampling times

which could be in the present study of the type 'lake A at time 1', 'lake B at time 5' etc. For getting separate PC-representations of amino acids, lakes and sampling times, a so-called N-way PCA is required which may be understood as a non-trivial extension of conventional PCA. It is based on a model by Tucker [34] and on a solution algorithm proposed by Kroonenberg and De Leeuw [35]. A tutorial on this topic can be found in [36]. Taking the data which the plots in Fig. 7 were based on, but separately evaluated there, one can arrange them in a three dimensional array corresponding to Fig. 8 with 4 rows (sampling times), 15 columns (amino acids) and 10 slices (lakes). The resulting N-way PC-plot is shown in Fig. 9. This plot may be considered as an extraction of the main data structure present in that part of the whole data collection with coinciding elements (constant set of lakes, amino acids and sampling times). It allows convenient interpretation of amino acids both with respect to lakes and to sampling times. Again, all amino acids get an equal sign along the horizontal axis making it interpretable as a general factor of amino acid production. All sampling times have practically equal weight on the horizontal axis, this factor is demonstrated to be stable over time. It distinguishes the lakes along their horizontal positions in the growing order Stechlinsee, Große Fuchskuhle (four parts) < Müggelsee < Werlsee, Peetzsee, Flakensee < Möllensee and in good accordance with their geographical positions (see Fig. 1). The close positions of the four parts of Große Fuchskuhle confirm a good reproducibility of experimental data from all amino acids over all sampling times. The vertical axis in the plot is a specific factor which reflects the already mentioned (lys, met, arg)-peak in April (compare the vertical positions of the respective points). It becomes clear from the picture that in the considered range of sampling times the major fraction of lakes is distinguished by the degree of the overall amino acid production. Lake Müggelsee is the only one to show a differing pattern which, however, is not stable over time. Therefore it remains an open question whether anthropogenic influences lead to patterns which are characteristic for different lakes. At least, there seem to exist reflections of short-term fluctuations. To understand the reasons of different amino acid patterns, especially for the lake Müggelsee, and their changes in time more detailed investigations are necessary, e.g. of amino acid patterns of different solutions of artificially cultivated different species of algae. Results will be reported later.

Acknowledgement. Thanks go to Dr. Behrendt and Dr. Koschel (Institut für Gewässerökologie und Binnenfischerei, Berlin and Neuglobsow) for providing water samples.

References

1. Schürmann J (1964) Vjschr Naturforsch Ges Zürich 109:409–460
2. Hellebust JA (1965) Limnol Oceanogr 10:192–206
3. Gocke K (1970) Arch Hydrobiol 67:285–367
4. Ittekkot V, Martins O, Seifert R (1983) Mitt Geol Paläontol Inst Univ Hamburg 55:119–127
5. Awapara J (1962) In: Holten JT (ed) Amino acid pools. Elsevier, Amsterdam, pp 158–205
6. Henrichs SM, Cuhel R (1985) Appl Environm Microbiol 50: 543–545
7. King K, Hare PE (1972) Micropaleontology 18:285–293
8. Punett T, Derrenbacker EC (1966) J Gen Microbiol 44:114–150
9. Gardner WS, Lee GF (1978) Mar Chem 6:27–40
10. Palmork KH (1963) Acta Chem Scand 17:1456
11. Siegel A, Degens ET (1966) Science 151:1098–1101
12. Moore SW, Stein H (1954) J Biol Chem 211:907–913
13. Roth SE (1992) Dissertation, Engler-Bunte-Institut, Karlsruhe
14. Roth SE, Maier D (1989) Vom Wasser 73:303–314
15. Roth SE, Maier D, Beran D (1991) Vom Wasser 76:61–71
16. Günther S (1992) Master Theses, Humboldt University, Berlin
17. Jorgensen NOG, Sondergaard M, Hansen HJ, Bosselmann S, Rieman R (1983) Hydrobiolog 107:107–122
18. Jorgensen NOG (1986) Feshwater Biol 16:255–268
19. Jorgensen NOG (1987) Limnol Oceanogr 32:97–111
20. Einarsson S (1985) J Chromatogr 348:213–216
21. Cohen SA, Tarvin TL, Bidlingmeyer BA (1984) J Chromatogr 333:93–104
22. Garrasi C, Degens ET, Mopper K (1979) Mar Chem 8:71–85
23. Cox KE, Claibome FB (1941) J Amer Water Works Ass 41: 948–952
24. Hellwig DHR (1964) Intern J Air Water Poll 8:215–228
25. Funk W (1977) Vom Wasser 48:75–87
26. Henrion R, Henrion G (1994) Multivariate Datenanalyse. Springer, Berlin Heidelberg New York
27. KNAUER Wissenschaftliche Gerätebau GmbH (1990) The new KNAUER Amino Acid Analyzer – a versatile system. Manual
28. Michael G, Nass D, Henrion G (in preparation)
29. Johnson RA, Wichern DW (1982) Applied multivariate statistical analysis. Prentice Hall, New Jersey
30. Sharaf MA, Illmann DL, Kowalski BR (1986) Chemometrics. Wiley, New York
31. Hoeg S (1983) Acta Hydrophys 28:5–36
32. Driescher E, Behrendt H, Schellenberger G, Stellmacher R (1993) Int Rev Ges Hydrobiol 78:327–343
33. Geladi P, Isaksson H, Lindqvist L, Wold S, Esbensen K (1989) Chemom Intell Lab Syst 5:209–220
34. Tucker LR (1966) Psychometrika 31:279–311
35. Kroonenberg PM, De Leeuw J (1980) Psychometrika 45:69–97
36. Henrion R (1994) Chemom Intell Lab Syst 25:1–23