Not to be cited without
permission of the authors[1]

Canadian Atlantic Fisheries
Scientific Advisory Committee

CAFSAC Research Document 91/9

Ne pas citer sans
autorisation des auteurs[1]

Comité scientifique consultatif des
pêches canadiennes dans l'Atlantique

CSCPCA Document de recherche 91/ 9

Improving the performance of classification and composition methodology for salmon of mixed continental origin at West Greenland

by

R. B. Millar and D. G. Reddin
Science Branch
Department of Fisheries and Oceans
P. O. Box 5667
St. John's, Newfoundland A1C 5X1

[1] This series documents the scientific basis for fisheries management advice in Atlantic Canada. As such, it addresses the issues of the day in the time frames required and the Research Documents it contains are not intended as definitive statements on the subjects addressed but rather as progress reports on ongoing investigations.

Research Documents are produced in the official language in which they are provided to the Secretariat by the author.

[1] Cette série documente les bases scientifiques des conseils de gestion des pêches sur la côte atlantique du Canada. Comme telle, elle couvre les problèmes actuels selon les échéanciers voulus et les Documents de recherche qu'elle contient ne doivent pas être considérés comme des énoncés finals sur les sujets traités mais plutôt comme des rapports d'étape sur les études en cours.

Les Documents de recherche sont publiés dans la langue officielle utilisée par les auteurs dans le manuscrit envoyé au secrétariat.

ABSTRACT

We investigated the effect on classification accuracy of stratifying by river age and compared the performance of the classification correction and maximum likelihood estimators of composition. Scale data on known origin salmon for years 82,85,86,87,88 and 89 were used. Scatter plots of the two measured scale variables were produced and based on these plots we chose to look at years 85 and 89 in more detail. Stratifying by river age degraded the 1985 classification rates, but produced a slight improvement for the 1989 data. Simulation studies were performed to show the effect on composition estimation accuracy. For small mixed fishery samples the maximum likelihood estimator performed better, but for large mixed fishery samples the classification correction estimator was better.

RÉSUMÉ

Les auteurs ont étudié les effets, sur l'exactitude de la classification, de la stratification par temps passé en eau douce et comparé le rendement des estimateurs de correction de classification et de maximum de vraisemblance de composition. L'étude a porté sur des données scalimétriques de saumons d'origine connue obtenues pour les années 1982, 1985, et 1986 à 1989. Des diagrammes de dispersion obtenus pour les deux variables scalimétriques déterminées ont permis de choisir les années 1985 et 1989 pour un examen plus détaillé. La stratification par le temps passé en eau douce a dégradé les taux de classification de 1985 mais a permis d'obtenir une légère amélioration avec les données de 1989. Des simulations ont été effectuées afin de déterminer les effets sur l'exactitude de l'estimation de la composition. Dans le cas de petits échantillons mixtes de pêche, l'estimateur de maximum de vraisemblance a donné de meilleurs résultats, tandis que l'estimateur de correction de classification s'est avéré préférable pour les grands échantillons mixtes de pêche.

INTRODUCTION

Classification and composition estimation of salmon at West Greenland has historically used two scale circuli count variables CS1S and CS1W (Reddin 1986), and more recently only CS1S. Composition estimation has been implemented using the classification correction procedure (Cook and Lord 1978; Pella and Robertson 1979).

This study examines whether stratifying by river age provides an improvement in classification rates (and hence an improvement in the classification correction estimators of composition). In addition, the performance of the classification correction estimators is compared to that of the maximum likelihood estimator of composition (Millar 1990).

METHODS

Figures 1-6 show the scale variables CS1S and CS1W for years 82,85,86,87,88 and 89 for fish of known origin. The upper two plots on each figure show the entire data sets for North American and European origin fish. The bottom eight plots show these data partitioned by river age. All plots use the same horizontal and vertical scales.

In some years it is evident that the separation between North American and European fish is dependent on river age. For example, In 1985 there appears to be almost perfect separation between the river age 1 fish. Also, the means of the variables CS1S and CS1W may be quite different between different river ages. In 1985 it looks like the mean of CS1S decreases with increased river age. Thus, one might hope for improvements in classification accuracy if the known origin data is stratified by river age.

In the work described below it has been assumed that regardless of origin and river age, the variables CS1S and CS1W are approximately normally distributed and have a common covariance matrix. Then linear discriminant analysis is appropriate. From Figures 1-6 it is evident that the assumption of a common covariance matrix is questionable. The use of different covariance matrices (quadratic discriminant analysis) is something that we intend to explore next.

RESULTS

Tables 1a shows the classification matrices for 1985 and 1989 for both unstratified and stratified (by river age) linear discriminant analyses. The 1985 data contains 105 and 132 fish of North American and European origin respectively. The numbers are 170 and 27 for 1989. Since there are only two 1985 European river age 4 fish, river age 4 was combined with river age 3 for the 1985 analyses. Similarly, river ages 3 and 4 were combined with river age 2 for 1989. Table 1b shows the classication broken down by river

age. As expected, the two origins of river age 1 fish are well separated in 1985 and there is no misclassification.

Classification rates are degraded by river age stratification in 1985. The classification of North American fish is unchanged but 5 more of the European fish are misclassified by the stratified approach (Table 1a). The effect of this is manifested in the classification correction estimates of composition (Tables 2a and 2b) and it is seen that the mean squared error increases under stratification. The 1989 data behaves in the opposite manner. Classification rates and composition estimation both improve.

The difference in mean squared error of composition estimation between the unstratified and stratified methods is modest. The effect of stratification is to increase mse by up to 20% in 1985 (fixed learning sample run) and to decrease mse by up to 35% in 1989 (random learning sample run).

Stratification has little affect on the maximum likelihood estimator of composition. This is to be expected since, for example, maximum likelihood implicitly realizes the separation between the river age 1 fish in 1985 by the simple fact that the North American river age 1 fish are not typical of European fish.

A comparison of Tables 2a and 2b shows that maximum likelihood does better than classification correction for mixed sample size 100 but worse for mixed sample size 400. This is because the maximum likelihood method exhibits a systematic bias (due primarily to the failure of the assumption of a common covariance matrix) which does not decrease with increasing sample size. Tables 3a and 3b break the mean squared error values from Tables 2a and 2b down into their bias and variability components. For small sample sizes random variability is the major contributor to the mean squared error and maximum likelihood outperforms classification correction. However, for large sample sizes the random variability is reduced but the maximum likelihood estimators bias is not, and so the classification correction is better.

DISCUSSION

The effect of river age stratification is extremely variable between years. It gave worse performance in 1985 but better performance in 1989. However, there is nothing to lose by calculating the stratified classification matrix. If it has worse classification rates than the unstratified then stratification is not used. As in the case of 1989, if the stratified classification rates are better then one can perform simulations to determine the expected gain in composition estimation. These simulations should include resampling the known origin fish to insure that the improvement is genuine, and is not just due to the extra parameters used in the discrimination procedure (stratification by river age requires

estimating the means of the scale variables for each river age within each origin, rather that just means for each origin in the unstratified case).

The maximum likelihood estimator of composition exhibited considerable bias, but considerably less variability than the classification correction estimator.. This can probably be foretold by the non-symmetry of the classification matrices. In 1985 there was more misclassification to European than to North American and maximum likelihood overestimated the European contribution. In 1989 the situation was reversed - there was more misclassification to North American than to European and maximum likelihood overestimated the North American contribution.

The non-symmetric classification matrix indicates that the assumption of a common covariance matrix is not valid. Using different covariance matrices (quadratic discriminant analysis) will hopefully produce more symmetry in the classification matrix. Also, since the bias of the maximum likelihood estimator is systematic, it can be reduced (Beacham et al. 1985). At this stage we feel that a little more work needs to be done before making a final comparison between the classification correction and maximum likelihood estimators of composition.

REFERENCES

Cook, R. C., and G. E. Lord. 1978. Identification of stocks of Bristol Bay sockeye salmon, *Oncorhynchus nerka,* by evaluating scale patterns with a polynomial discriminant method. Fish. Bull. 76: 415-423.

Beachan, T. D., R. E. Withler, and Allan P. Gould. 1985. Biochemical genetic stock identification of chum salmon (*Oncorhynchus keta*) in Southern British Columbia. Can. J. Fish. Aquat. Sci. 42: 437-448.

Millar, R. B. 1990. Comparison of methods for estimating mixed stock fishery composition. Can. J. Fish. Aquat. Sci. 47: 2235-2241.

Pella, J. J., and T. L. Robertson. 1979. Assessment of composition of stock mixtures. Fish. Bull. 77: 387-398.

Reddin, D. G. 1986. Discrimination between Atlantic salmon (*Salmo salar* L.) of North American and European origin. J. Cons. int. Explor. Mer. 43: 50-58.

Table 1a. Classification matrices for the 1985 and 1989 data for unstratified and stratified (by river age) classification.

1985

Unstratified

	from NA	from Euro
to NA	66	25
to Euro	39	107

	from NA	from Euro
to NA	.629	.189
to Euro	.371	.811

Stratified

	from NA	from Euro
to NA	66	30
to Euro	39	102

	from NA	from Euro
to NA	.629	.227
to Euro	.371	.773

1989

Unstratified

	from NA	from Euro
to NA	138	10
to Euro	32	17

	from NA	from Euro
to NA	.812	.370
to Euro	.188	.630

Stratified

	from NA	from Euro
to NA	144	10
to Euro	26	17

	from NA	from Euro
to NA	.847	.370
to Euro	.153	.630

Table 1b. Classification broken down by river age.

1985

	from NA1	from NA2	from NA3	from E1	from E2	from E3
to NA1	8	0	0	0	0	0
to NA2	0	20	0	0	23	0
to NA3	0	0	38	0	0	7
to E1	0	0	0	21	0	0
to E2	0	10	0	0	56	0
to E3	0	0	29	0	0	25

1989

	from NA1	from NA2	from E1	from E2
to NA1	71	0	6	0
to NA2	0	73	0	4
to E1	8	0	9	0
to E2	0	18	0	8

Table 2a. Mean squared errors for estimation of the continental composition of Atlantic salmon at West Greenland. One hundred simulations were performed in each run. Size of mixed fishery sample is 100.

1985

	Classification correction		Maximum likelihood	
	unstratified	stratified	unstratified	stratified
fixed learning sample	.0146	.0183	.0093	.0103
random learning sample	.0211	.0234	.0130	.0134

1989

	Classification correction		Maximum likelihood	
	unstratified	stratified	unstratified	stratified
fixed learning sample	.0153	.0130	.0117	.0130
random learning sample	.0307	.0228	.0228	.0217

Table 2b. Mean squared errors for estimation of the continental composition of Atlantic salmon at West Greenland. One hundred simulations were performed in each run. Size of mixed fishery sample is 400.

1985

	Classification correction		Maximum likelihood	
	unstratified	stratified	unstratified	stratified
fixed learning sample	.0028	.0039	.0047	.0046
random learning sample	.0078	.0081	.0068	.0071

1989

	Classification correction		Maximum likelihood	
	unstratified	stratified	unstratified	stratified
fixed learning sample	.0031	.0024	.0071	.0085
random learning sample	.0149	.0108	.0123	.0156

Table 3a. Mean and standard deviation of estimated North American contribution of Atlantic salmon at West Greenland. One hundred simulations were performed in each run. Each simulation constructed a mixed fishery sample of size 100 from a mixed fishery with true composition (0.5,0.5).

1985

	Classification correction		Maximum likelihood	
	unstratified	stratified	unstratified	stratified
fixed				
learning	.515	.518	.466	.453
sample	.120	.134	.090	.090
random				
learning	.512	.517	.464	.448
sample	.145	.152	.108	.103

1989

	Classification correction		Maximum likelihood	
	unstratified	stratified	unstratified	stratified
fixed				
learning	.500	.500	.570	.579
sample	.124	.114	.082	.082
random				
learning	.491	.516	.556	.576
sample	.175	.150	.141	.126

Table 3b. Mean and standard deviation of estimated North American contribution of Atlantic salmon at West Greenland. One hundred simulations were performed in each run. Each simulation constructed a mixed fishery sample of size 400 from a mixed fishery with true composition (0.5,0.5).

1985

	Classification correction		Maximum likelihood	
	unstratified	stratified	unstratified	stratified
fixed				
learning	.493	.502	.448	.445
sample	.052	.062	.045	.040
random				
learning	.485	.502	.459	.446
sample	.087	.090	.072	.065

1989

	Classification correction		Maximum likelihood	
	unstratified	stratified	unstratified	stratified
fixed				
learning	.502	.503	.574	.585
sample	.056	.049	.039	.037
random				
learning	.492	.545	.578	.604
sample	.122	.094	.078	.069

Scatter plots of CS1W vs CS1S for combined and individual river ages, 1982

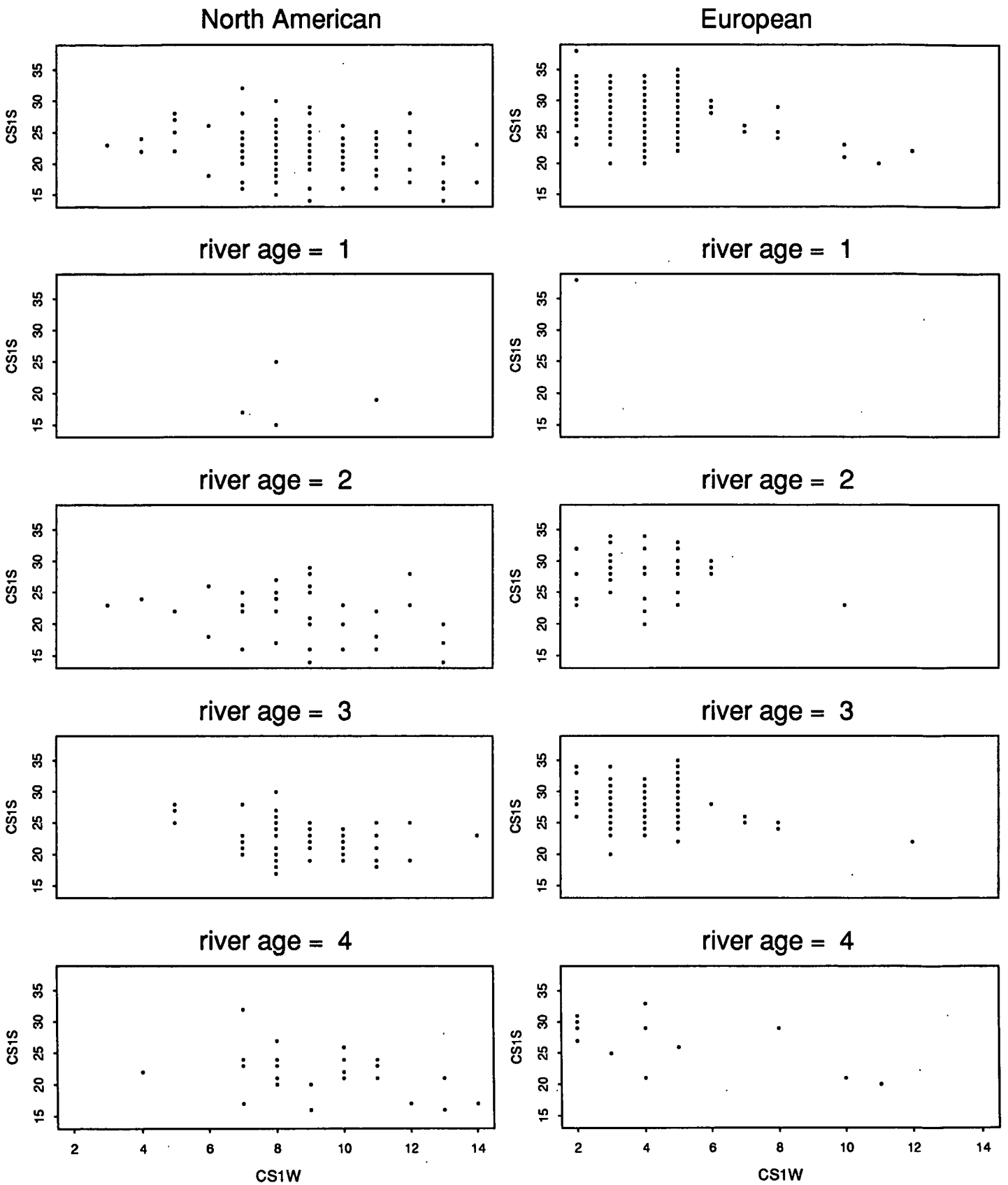

Figure 1

Scatter plots of CS1W vs CS1S for combined and individual river ages, 1985

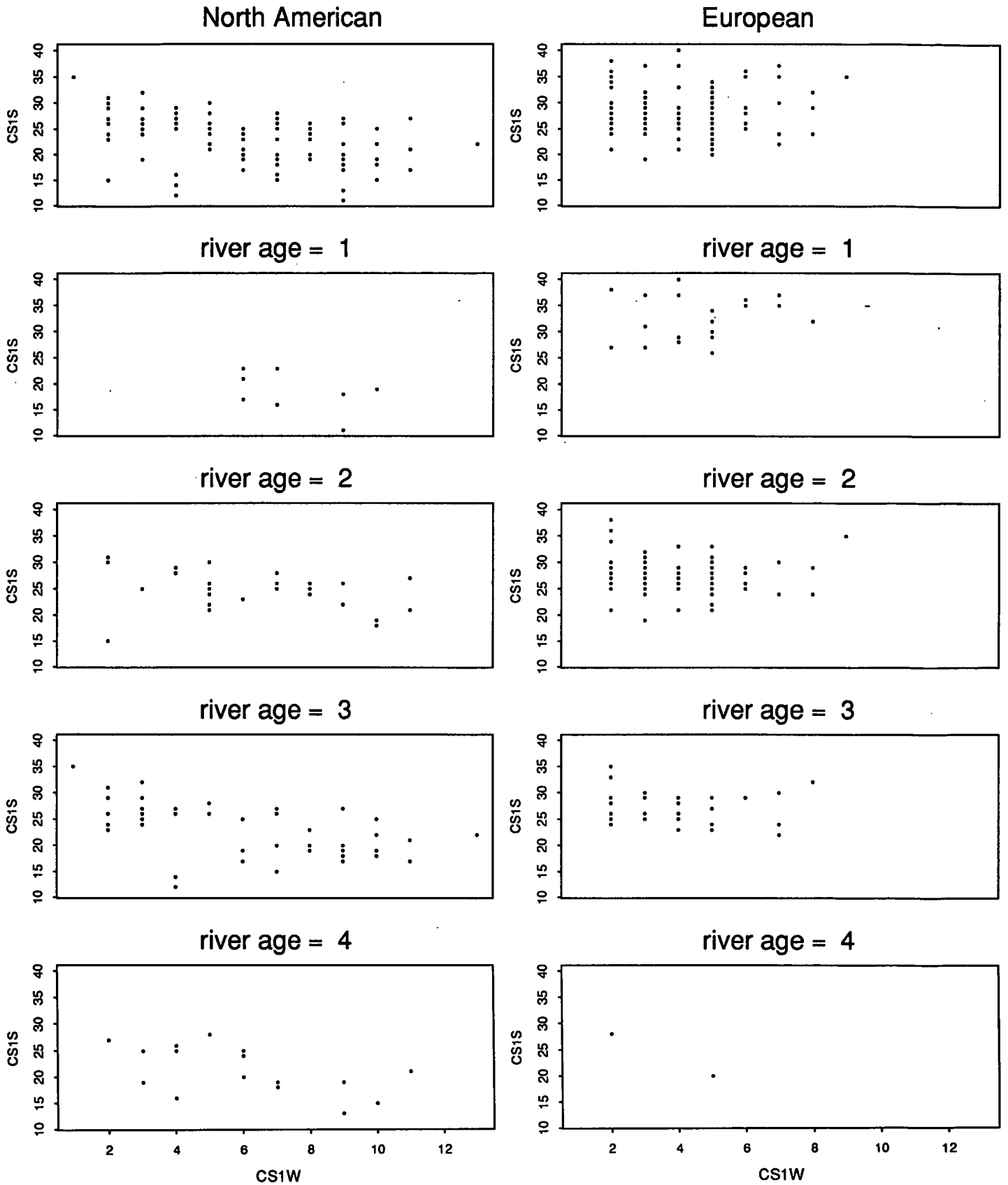

Figure 2

Scatter plots of CS1W vs CS1S for combined and individual river ages, 1986

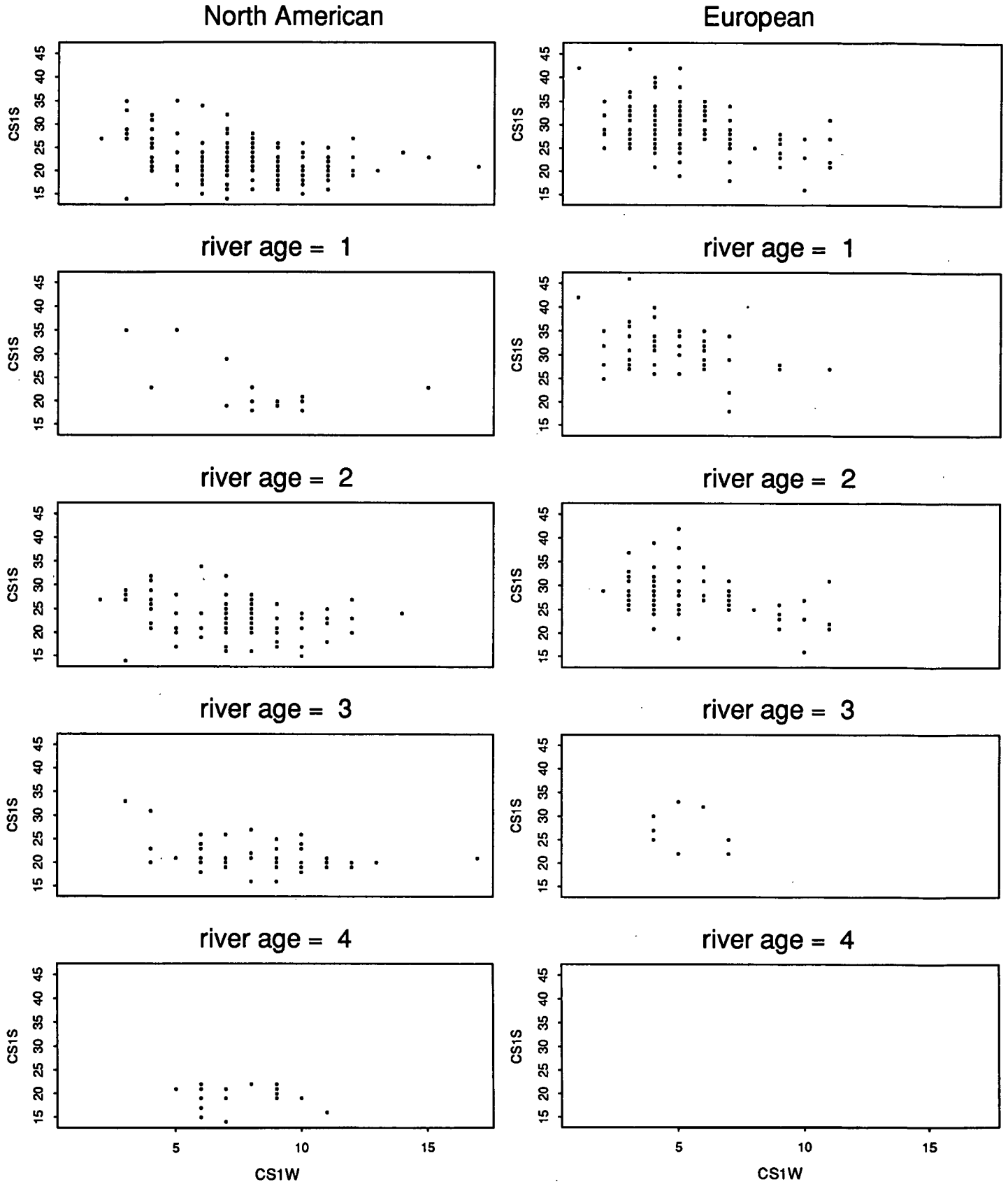

Figure 3

Scatter plots of CS1W vs CS1S for combined and individual river ages, 1987

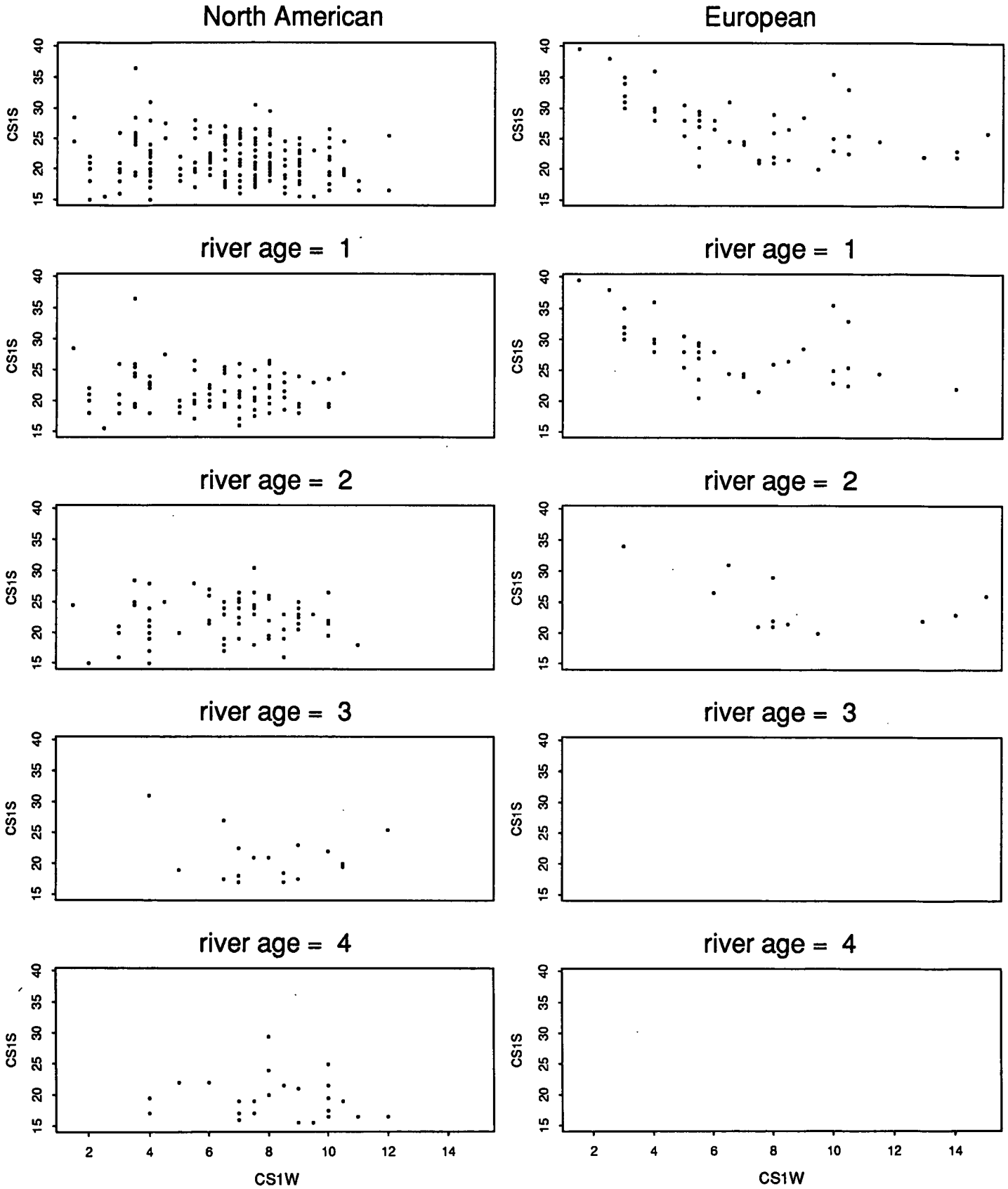

Figure 4

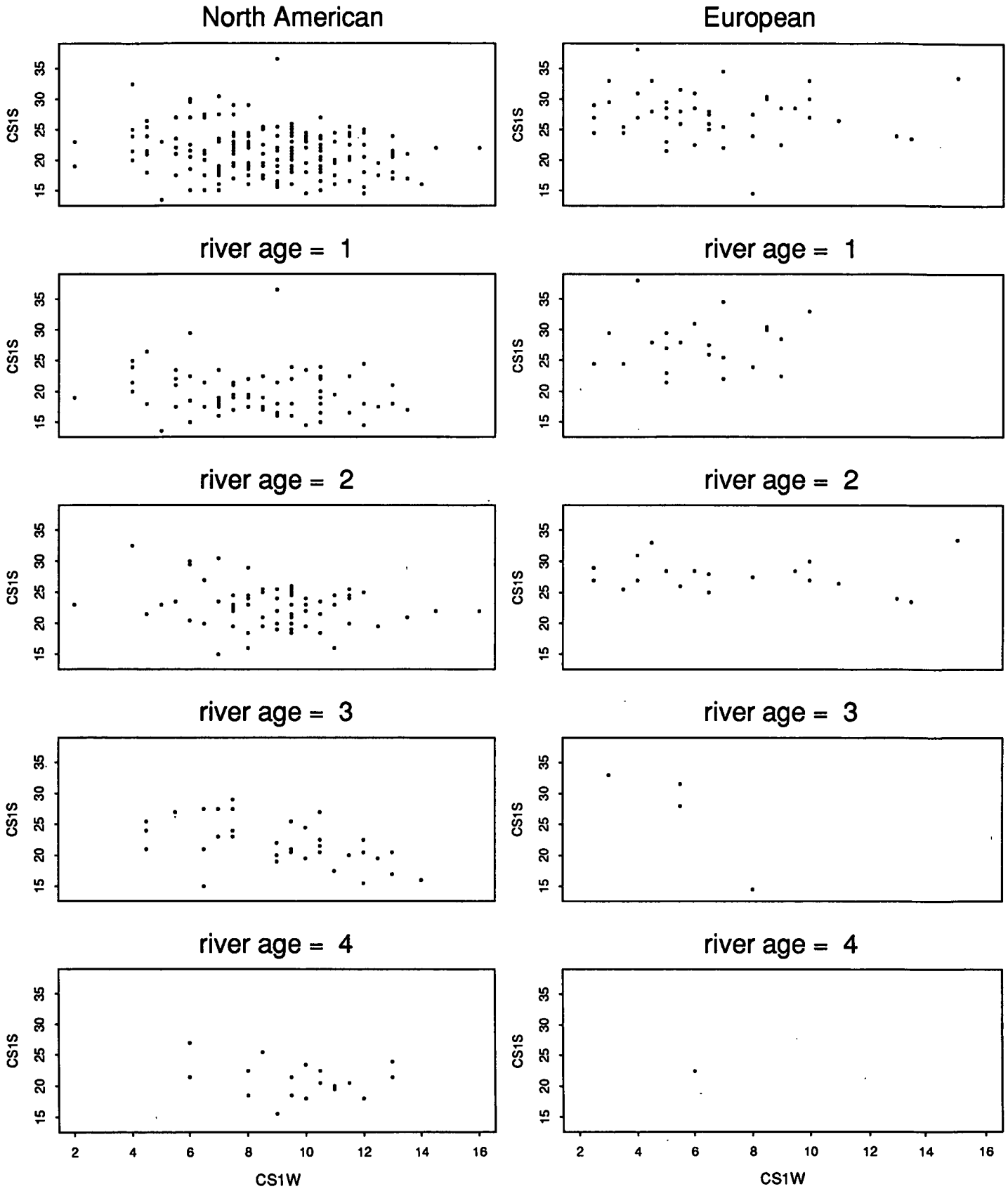
Scatter plots of CS1W vs CS1S for combined and individual river ages, 1988
North American
European
river age = 1
river age = 1
river age = 2
river age = 2
river age = 3
river age = 3
river age = 4
river age = 4
CS1S
CS1W
15
20
25
30
35
2
4
6
8
10
12
14
16

Figure 5

Scatter plots of CS1W vs CS1S for combined and individual river ages, 1989

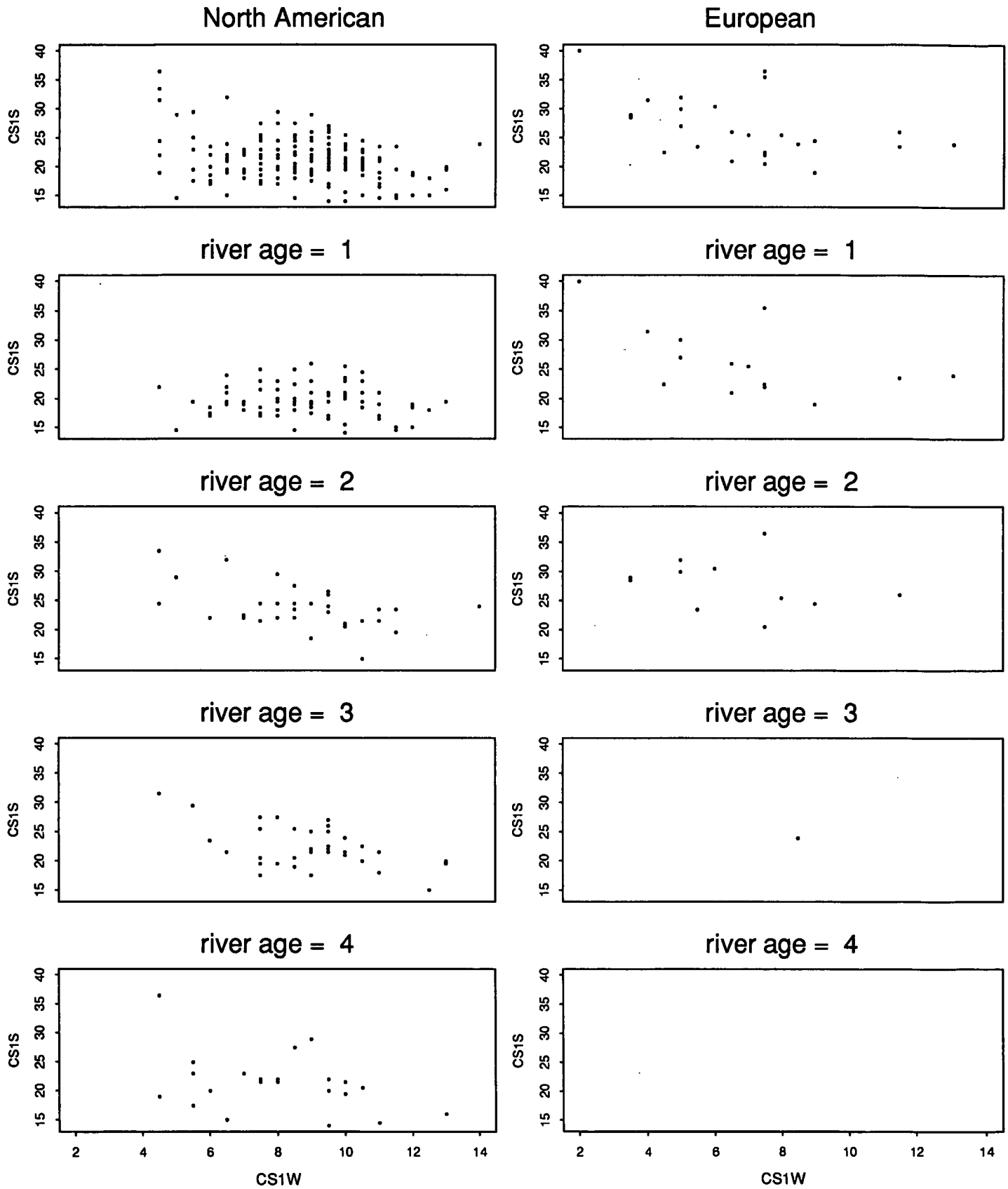

Figure 6