NPAFC Doc. 375

Rev.

The precision and accuracy of genetic baseline data sets for pink, chum and sockeye salmon stock identification in Pacific Ocean mixed-fisheries catches

by

Natalia V. Varnavskaya
KamchatNIRO
Petropavlovsk-Kamchatsky 683000
Naberejnaya 18

submitted to the

North Pacific Anadromous Fish Commission

by
Russia

November 1998

THIS PAPER MAY BE CITED IN THE FOLLOWING MANNER:
N. V. Varnavskaya. 1998. The precision and accuracy of genetic baseline data sets for pink, chum and sockeye salmon stock identification in Pacific Ocean mixed-fisheries catches. (NPAFC Doc. 375). 16p

Introduction

Electrophoretic studies of genetic variation at enzyme loci developed last several years have allowed to create large data sets of allelic frequencies for many populations of Pacific salmon from the Pacific Rim rivers and lakes. The statistical analysis of those data sets has shown significant regional heterogeneity which could be used for stock identification in mixed catches in the Pacific Ocean and adjusted sea. (Winans et al., 1994; Wilmot et a., 1994; Varnavskaya et al, 1994a,b; Wood et al., 1994; Kondzela et al., 1994, Varnavskaya et al., 1996). The important application of mixed-stock fisheries identification data is to reveal the migration routes of Pacific salmon of different origin in the Pacific Ocean. This is necessary for understanding the distribution and ecology of local stocks and regional groups of stocks of salmon during their sea life period. In preliminary publications the importance of selection of sets of loci used for analysis was shown (Hawkins et al, 1996).

The purpose of this study was to create the data sets from selected loci and populations, and to examine their ability to be used as a tool in separating mixed-stock fisheries chum, sockeye, and pink catches by testing the hypothetical mixed fisheries samples with known composition using the Conditional Maximum Likelihood Estimator (MLE).

Material and Methods

The data sets (Table 1) were selected from baselines collected in Northwest Fisheries Center, Seattle, Washington, USA, Auke Bay

Laboratory, Juneau, Alaska, USA, KamchatNIRO Genetic Laboratory, Petropavlovsk-Kamchatsky, Russia, Pacific Biological Station, Nanaimo, Canada (Winans et al., 1994; Wilmot et a., 1994; Varnavskaya et al, 1994a,b; Wood et al., 1994; Kondzela et al., in press, Varnavskaya et al., 1996; Noll et al., in press).

Table 1. The characterization of genetic data sets used for analysis.

Species	Number of loci	Number of populations
Chum	48	40
Sockeye	8	61
Pink, even	38	26
Pink, odd	31	19

The chum baseline included only Asian populations, the pink baseline included Asian and some of Alaskan populations, and the sockeye baseline included selected populations through all the Pacific Rim area.

A series of hypothetical mixtures were simulated from the baselines using the program SIMULATR developed by Auke Bay Laboratory (Masuda et al. 1991) The random samples from the baseline data were generated, and mixtures of specified size and composition were created. Each simulated mixture was then analyzed using the Conditional Maximum Likelihood Estimate (MLE) method (Pella and Milner, 1987). There were two scenarios for simulations. An equal percentage of abundance from each of stocks was used in all simulated mixtures (simulation 1). Series of estimates were

computed with mixtures composed of 100% of each region (simulation 2). The mean estimates, and standard deviations were calculated for standardized comparison by each region.

Results and discussion

The ability of genetic baseline to be used as a tool for mixed-stock identification procedure depends on two important characteristics: observed genetic diversity between baseline populations should be enough significant to provide the precision of estimates, and sample size of a sea mixture should be enough big to provide the accuracy of estimates. To be successful in achieving the first goal we need to select the number of loci which significantly differ in populations. If the second level of hierarchy needs to be used, we must make sure it was chosen in accordance to real hierarchical structure revealed by genetic diversity analysis. It means that genetic distance coefficients should be approximately equal at each level. If this condition was not provided the genetic similarity in some pairs of regional groups would fail the whole set of estimates. We divided baseline data sets in regional groups in accordance with traditional way used in fisheries management of different countries, and performed the MLE-analysis of simulated mixtures of sample size 100 fish each (Tables 2-5).

In pink salmon both odd and even-year broodlines the differences between calculated and expected estimates were insignificant when the simulation 1-scenario was used. Results for the "worst case scenario", where the actual contribution approaches the

limit of the MLE, 0 or 100%, the average estimates showed that in even-year pink salmon the Hokkaido population could be identify with precision 85% and about 9% could be mistaken with Sakhalin populations. The same was for the Northwest and West Kamchatka. The Iturup and Sakhalin populations had even more similarity, so the overlap could be about 15%. All other regions could be identified in mixtures with precision close to 90%.

The precision of odd-year pink baseline with 31 loci allelic frequencies data in 19 populations was close to 90% for all regional groups, except of Magadan and East Kamchatka populations which were similar to each other. The overlap between them was about 15%, and with the West Kamchatka region it was about 5% each.

To test the chum salmon baseline we created more regional groups, so the stocks from large river basins were distinguished, such as the Amur River, Penjina River, Kamchatka River, Anadyr River. This baseline data set was able to provide separation with precision more than 90% for all regions except the East Kamchatka which was similar to the West Kamchatka, and the Penjina River which was similar to the West Kamchatka, and the Magadan region which was also similar to the West Kamchatka, so precision of identification of those three regions was about 80%.

The sockeye baseline data set included allelic frequencies for 8 high polymorphic loci in 61 populations from Asia and North America. The precision of this baseline is not so good, as for chum, but still it is able to identify almost all regions with precision about 80%. The Meziadin and Fraser River populations could be identify with precision

more than 95%. There is some similarity between the Kurilskoye Lake stock and other West Kamchatka stocks. The sockeye baseline is desirable to be complete with more polymorphic loci, and we plan that for our next studies.

We tested the influence of a sample size of a mixture on accuracy of MLE-estimations for salmon baselines. Each separation of a simulated mixture of 100% of one regional group was repeated 20 times with sample size from 50 to 1000 fish for chum and 10 times with sample size from 50 to 500 for sockeye and even pink salmon. In the most regions the standard deviations did not increase after sample size was 100 fish (Fig. 1, 2, 3). The mean deviations did not go any closer to 100%. That allow us to make a conclusion that for this particular chum salmon baseline data set the sample size about 100 fish is enough to make significant estimations. Even a 50 fish sample size could provide an information about stock composition but in this case the standard error would increase. The relationships between mixture sample sizes and coefficients of variation confirm this conclusion (Fig. 4). The composition estimates with sockeye and pink baselines would demand at least 100-200 fish in mixture sample.

Cited literature

1. Hawkins S., Varnavskaya N., V. Efremov, Wilmot R. L. Simulations of the even-year Asian pink salmon genetic baseline to determine accuracy and precision of stock composition estimates.

NPAFC Int. Symp. Asses. & Status of Pacific Rim Salm. Stocks, Sapporo, 1996. P. 34.

2. Kondzela, C.M., C.M. Guthrie, S. L. Hawkins, C.D. Russell, J.H. Helle, and A.J. Gharrett. Genetic relationships among chum salmon population in southeast Alaska and northern British Columbia salmon // Can. J. Fish. Aquat. Sci. 1994. V. 51 (Suppl.). P. 50-64.
3. Masuda, M., S. Nelson and J. Pella. The computer program for computing conditional maximum likelihood estimates of stock composition from discrete characters. USA-DOC-NOAA-NMFS, Auke Bay Laboratory. Auke Bay. AK. 1991
4. Noll C., Varnavskaya N. V., Matzak E., Hawkins S. Kondzela C., Midanaya V. V., Katugin O., Russell C., G. S. Fesunova, N. M. Kinas, Guthrie III C., Mayama H. Yamazaki F., Garrett A. J. Genetic relationships between even-year pink salmon (Oncorhynchus gorbuscha) from Asia and Alaska. Can. J. Fish Aquat. Sci. 199x.
5. Pella, J.J., and G. B. Milner. Use of genetic marks in stock composition analysis // In N. Ryman and F.Utter [ed.] Populations genetics and fishery management. University of Washington Press, Seattle, WA. 1987. P. 247-276.
6. Wilmot R. L., Everett R. J., Spearman W. J., Baccus R., Varnavskaya N. V, Putivkin S. V. Genetic stock structure of Western Alaska chum salmon and a comparison with Russian Far East stocks/. Can. J. Fish Aquat. Sci. 1994. V. 51 (Suppl.). P.84-94.

7. Winans G. A., Aebersold P. B., Urawa S., Varnavskaya N. V. Determinig continent of origin of chum salmon (Oncorhynchus keta) using genetic stock identification techniques: status of allozyme baseline on Asia. Can. J. Fish Aquat. Sci.1994. V.51 (Suppl.). P. 95-113.
8. Varnavskaya N. V., Wood C. C., Everett R. J. Genetic variation in sockeye salmon (Oncorhynchus nerka) populations of Asia and North America. Can. J. Fish Aquat. Sci. 1994. V. 51 (Suppl.). P. 132-146.
9. Varnavskaya N. V., Wood C. C., Everett R. J., Wilmot R. L., Varnavsky V. S., Midanaya V. V. , Quinn T. P. Genetic differentiation of subpopulations of sockeye salmon (Oncorhynchus nerka) within lakes of Alaska, British Columbia and Kamchatka. Can. J. Fish Aquat. Sci. 1994. V. 51 (Suppl.). P. 147-157.
10. Varnavskaya, N. V., C. M. Kondzela, R. L. Wilmot, V. Efremov, Xi. Luan, V. A. Davydenko, E. A. Sboeva, C. M. Guthrie III. Genetic variation in Asian populations of chum salmon, Oncorhynchus keta (Walbaum). NPAFC Int. Symp. Asses. & Status of Pacific Rim Salm. Stocks, Sapporo, 1996. P. 65.
11. Wood, C. C., B. E. Riddell, D. T. Rutherford, and R. E. Withler. Biochemical genetic survey of sockeye salmon (*Oncorhynchus nerka*) in Canada. Proceeding of Intern. Symp. Genetics of Fish and Shellfish, Juneau, Alaska, 1993 // J. Fish. Aquat. Sci. 1994.

Table 2. Mean estimated proportions (%) of fish originating from main spawning regions in simulated mixed-stock fisheries samples based on allelic frequencies at 31 loci in 26 populations of even year pink salmon . In Simulation 1 each population was taken as equal portion in mixed-stock sample, sample size was 200 fish. In Simulation 2 each regional group of populations was taken as 100% in mixed-stock sample, sample size was 100 fish..

Spawning region	Simulation 1		Simulation 2															
	exp.	cal.	exp.	cal.	exp.	cal.	exp.	cal.	exp.	cal.	exp.	cal.	exp.	cal.	exp.	cal.	exp.	cal.
Hokkaido	12	10.9	100	83.9	0	4.8	0	5.3	0	1.0	0	0.9	0	0.0	0	0.5	0	1.6
Iturup	4	3.9	0	0.3	100	77.0	0	0.5	0	0.0	0	0.0	0	0.0	0	0.8	0	0.0
Sakhalin	24	26.1	0	9.1	0	16.0	100	87.8	0	1.9	0	2.2	0	6.7	0	0.6	0	2.3
West Kamchatka	20	19.1	0	0.9	0	1.8	0	0.1	100	87.9	0	12.8	0	4.6	0	0.0	0	1.8
Northwest Kamchatka	4	5.2	0	0.0	0	0.0	0	1.1	0	7.1	100	81.6	0	0.0	0	0.0	0	1.2
Northeast Kamchatka	12	11.0	0	5.2	0	0.0	0	3.6	0	1.0	0	0.0	100	85.9	0	1.6	0	0.0
Magadan	4	4.3	0	0.0	0	0.0	0	0.6	0	0.6	0	0.0	0	1.7	100	95.9	0	1.0
Alaska	20	19.4	0	0.6	0	0.4	0	1.0	0	0.2	0	2.5	0	1.0	0	0.6	100	92.0

Table 3. Mean estimated proportions (%) of fish originating from main spawning regions in simulated mixed-stock fisheries samples based on allelic frequencies at 31 loci in 19 populations of odd year pink salmon. In Simulation 1 each population was taken as equal portion in mixed-stock sample, sample size was 200 fish. In Simulation 2 each regional group of populations was taken as 100% in mixed-stock sample, sample size was 100 fish..

	Simulation 1		**Simulation 2**											
Spawning region			**Hokkaido**		**Sakhalin**		**Magadan Coast**		**West Kamchatka**		**East Kamchatka**		**Alaska**	
	cal.	exp.	cal.	exp.	cal.	exp.	cal.	exp.	cal.	exp.	cal.	exp.	cal.	exp.
Hokkaido	9.5	10	**92.1**	**100**	2.0		2.6		2.7		1.8		1.0	
Sakhalin	11.8	12	3.8		**94.6**	**100**	3.9		4.4		3.4		2.0	
Magadan Coast	11.4	9	0.7		0.4		**72.3**	**100**	3.3		10.5		1.0	
West Kamchatka	10.4	9	2.1		1.5		6.4		**92.9**	**100**	4.5		2.0	
East Kamchatka	26.0	30	0.4		0.2		12.8		3.6		**77.8**	**100**	1.2	
Alaska	30.4	30	0.6		1.2		1.7		3.2		1.8		**92.6**	**100**

Table 4. Mean estimated proportions (%) of fish originating from main spawning regions in simulated mixed-stock fisheries samples based on allelic frequencies at 48 loci in 40 populations of Asian chum salmon. Simulated mixtures were composed of 100% of each region.

Spawning region	Simulation 2																			
	cal.	exp.	cal.	exp.	cal.	exp.	cal.	exp.	cal.	exp.	cal.	exp.	cal.	exp.	cal.	exp.	cal.	exp.	cal.	exp.
Anadyr River	92.3	100	1.1		1.3		3.5		0.2		0		0.4		0		0		0	
East Kamchatka	0.0		79.2	100	1.8		2.7		1.0		0		1		0.1		0		0	
Kamchatka River	0.8		2.2		90.6	100	0		0		0		0.5		0.4		0		0	
West Kamchatka	5.0		10.9		2.3		91.0	100	5.2		8.5		10.9		0		0		0.5	
Northwest Kamchatka	0.0		2.3		0		0		93.3	100	1.6		1.4		0		0		0	
Penjina River	0.0		0		0		0		0		84.1	100	0		0		0		0	
Magadan Coast	0.7		3.5		1.3		0.6		0		1.4		82.9	100	0.4		0		0.4	
Sakhalin Island	0.4		0.2		0.7		0		0		0.5		0.1		96.9	100	0.4		0	
Prymorye Coast	0.1		0.2		0.3		0.3		0		0.8		0		0		97.7	100	0	
Amur River	0.2		0		0.3		0.2		0.2		0.1		0.7		0		0.3		98.4	100
Japan	0.6		0		0.7		1.3		0		2.3		1		0.2		1.0		0.4	

Table 5. Mean estimated proportions (%) of fish originating from main spawning regions in simulated mixed-stock fisheries samples based on allelic frequencies at 8 loci in 61 populations of sockeye salmon. Simulated mixtures were composed of 100% of each region.

Spawning region	Simulation 2																			
	cal.	exp.	cal.	exp.	cal.	exp.	cal.	exp.	cal.	exp.	cal.	exp.	cal.	exp.	cal.	exp.	cal.	exp.	cal.	exp.
Northwest Kamchatka	71.6	100	9.6		6.0		5.0		0.0		4.1		0.7		2.2		0		0.9	
West Kamchatka	6.6		79.7	100	5.2		3.1		0.5		1.9		0.0		0.7		0		0.9	
Kurilskiye Lake	4.2		3.1		76.7	100	1.8		3.4		2.9		0.5		2.8		2.2		2.8	
Southeast Kamchatka	1.4		0.0		2.5		93.9	100	0.0		0.4		0.0		0.9		0		0.4	
Kamchatka River	0.0		0.0		1.7		0.2		83.2	100	8.5		1.4		3.0		0		1.5	
Iliamna Lake	0.0		0.7		5.7		3.3		2.9		78.1	100	0.5		1.1		8.1		6.2	
Southeast Alaska	0.0		0.0		0.0		0.0		2.4		0.0		93.4	100	0.1		1.3		4.0	
Meziadin Lake	0.0		0.0		0.7		0.4		0.4		0.0		0.7		97.7	100	0		0.0	
Babin Lake	0.0		0.0		4.4		4.9		0.4		8.0		0.7		0.0		83.0	100	10.6	
Fraser River	0.0		0.0		0.1		0.0		0.0		0.2		0.4		0.0		5.1		99.3	100

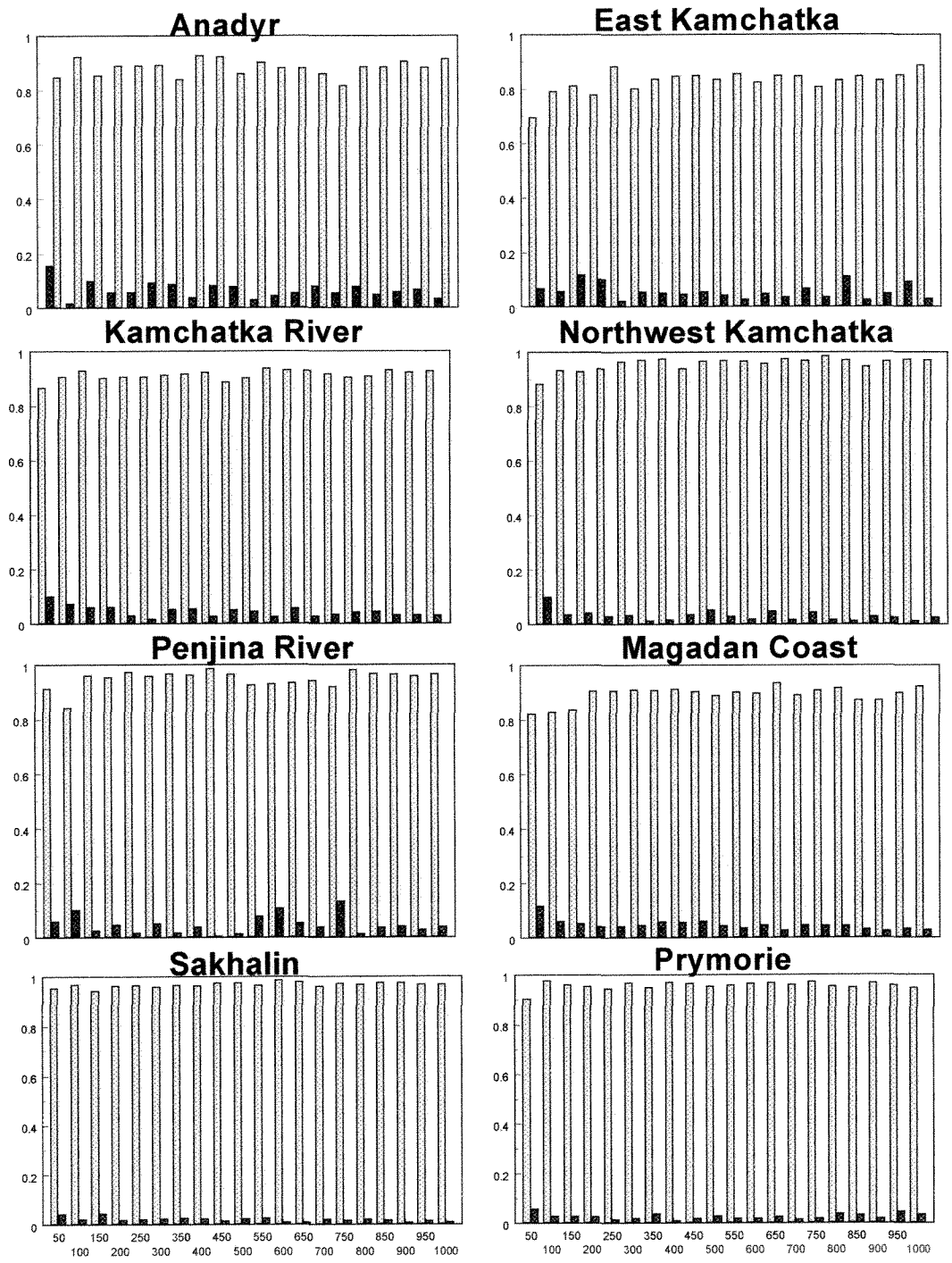


Figure 1. Mean estimates (light bars) and standart deviations (dark bars) for different mixture sample size (20 simulations from 50 to 1000 fish in mixture) when simulated mixtures were composed of 100% of each region for the database set of 48 polymorphic loci and 40 populations in chum salmon of the Pacific Rim.

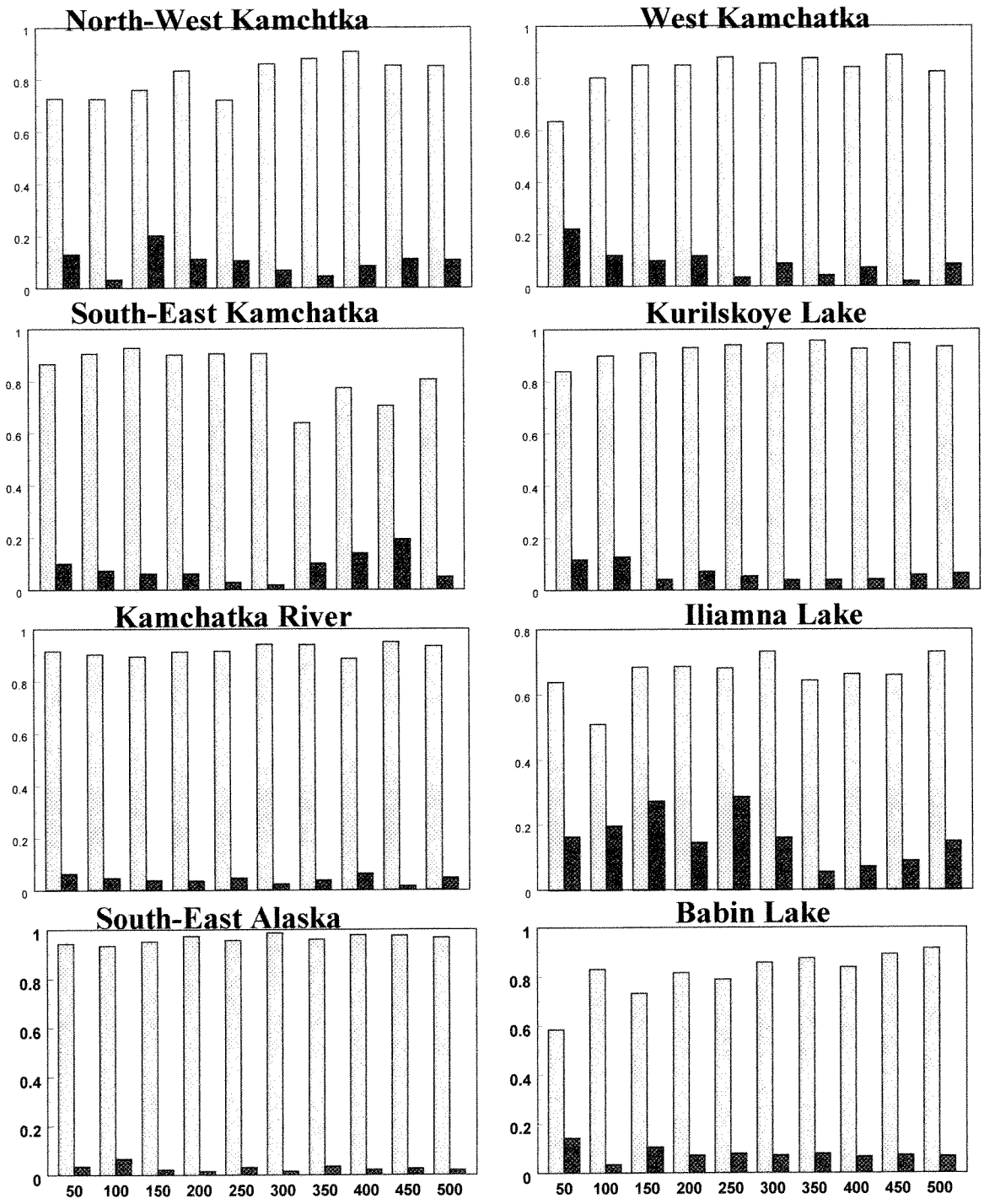


Figure 2. Mean estimates (light bars) and standard deviations (dark bars) for the different mixture sample size (10 simulations from 50 to 500 fish in mixture) when simulated mixtures were composed of 100% of each region for the database set of 8 high polymorphic loci and 61 populations in sockeye salmon of the Pacific Rim.

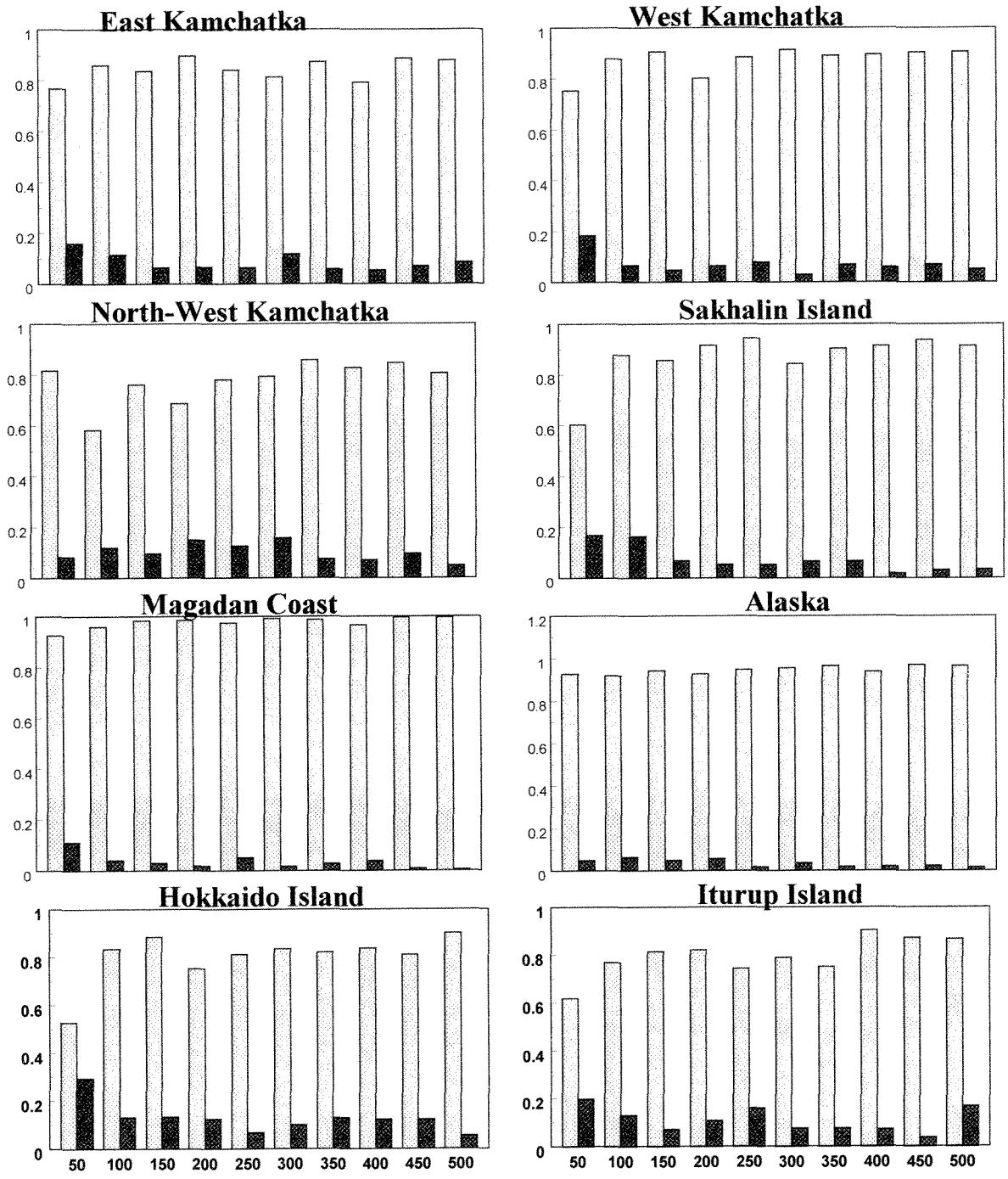
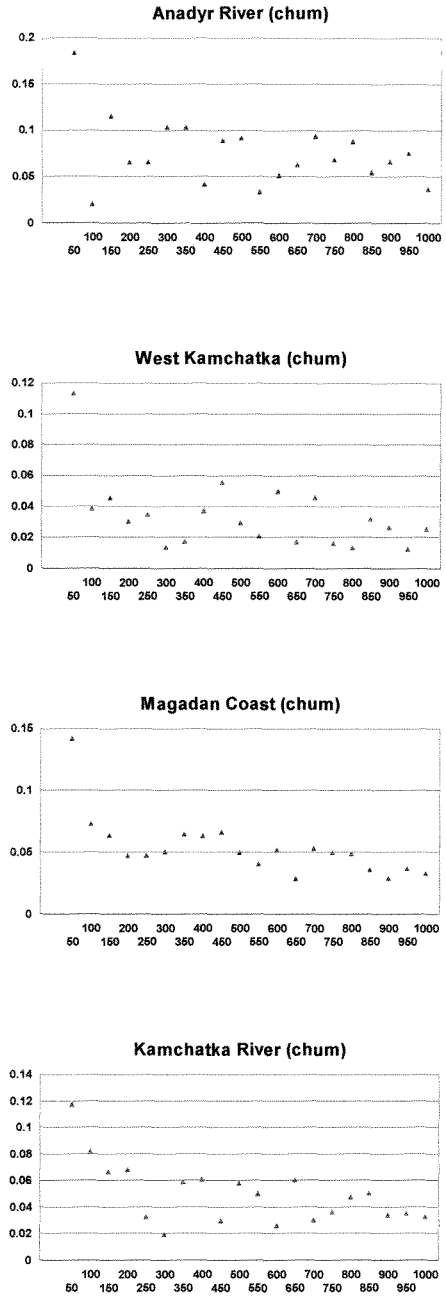


Figure 3. Mean estimates (light bars) and standard deviations (dark bars) for the different mixture sample size (10 simulations from 50 to 500 fish in mixture) when simulated mixtures were composed of 100% of each region for the database set of 38 polymorphic loci and 26 populations in even year pink salmon of the Pacific Rim.



1. Figure. 4. The relationship between sample size of mixture and the coefficient of variation. The simulations were made for the chum database set of 48 loci and 40 populations with 100% scenario for each of four regions.