

Further investigation of the difference in two datasets, raised by the Second CPUE modeling workshop, used for CPUE analyses of SBT

第2回CPUEモデリングワークショップで提起された、ミナミマグロのCPUE解析に使用するための2つのデータセットの違いに関する更なる検討

Hiroshi Shono and Tomoyuki Itoh
(庄野宏・伊藤智幸)

National Research Institute of Far Seas Fisheries, Fisheries Research Agency
(独立行政法人水産総合研究センター・遠洋水産研究所)

要約: 我々は、2007年の第2回CPUEワークショップでの解析に使用したショットバイショットデータから作成した2つのデータセットA、Bの違いについて検討した。CPUEトレンドは1990年代初め、特に1993年、1994年で違いが大きかった。ノミナルと標準化したCPUEを詳細に検討した結果、違いは4海区と9海区におけるCPUEが両データセットで違うことに起因すると思われた。ショットバイショットデータと5度×5度月集計データとで作成したデータにおいてノミナルCPUEにはほとんど違いがなかったことから、集計データによるさらなる解析は、ESCのメンバーの誰もが行い得る。標準化したCPUEの年トレンドは、GLMの説明変数とモデルの仮定にある程度依存する。簡単なバリデーションの結果、データセットBの方が、Aよりも統計学的に頑健だった。

Abstract: We mainly checked the difference of two datasets (Dataset-A and Dataset-B) made by shot-by-shot data used for CPUE analyses in the 2nd CPUE Workshop held in 2007. We obtained the different year trends of CPUE in the early 1990s, especially 1993-1994. As a result of investigation about nominal and standardized CPUE in detail, the difference seems to be mainly attributed the gap of CPUE values of area 4 and 9 in two datasets. Because there is little difference of nominal CPUE between the datasets made by shot-by-shot data and 5x5 month data, these and further investigation using the aggregated data (by 5x5/month) is also available for any of the ESC members. Year trend of standardized CPUE is dependent on the explanatory factors included into the GLM and its model assumption to some degree or another. Our result of a simple validation shows that the dataset-B is statistically more robust and stable than dataset-A.

Introduction

We mainly checked the difference of two datasets (Dataset-A and B) used for CPUE analyses of southern bluefin tuna from the viewpoints of nominal CPUE, standardized CPUE and statistical modeling for CPUE analyses etc. There datasets used in the 2nd CPUE workshop of CCSBT were defined as:

Dataset-A: Year(1992-2005), Area(4-9), Month(4-9) (past agreed definition)

Dataset-B: Year(1992-2005), Area(4,7,8,9), Month(Japanese fishing season)

Remark) Area and month defined in the dataset-B was annually changed

(See Table 4 of p.11, CCSBT-ESC/0709/SBT-Fisheries/Japan)

The two datasets is different from “LL1”, which is the past agreed and using in the annual calculation of CPUE indices regarding some points in Table A0.

Spatio-temporal coverage and Nominal CPUE

Spatio-temporal coverages of the datasets were compared in the number of hooks (Table A1). There were several area/month unique to either of the datasets. In addition to Area 5 and Area 6, Area 4 in April, July and August from 1991 to 1997, Area 7 in July from 1993 to 1996, and Area 9 in April, August and September from 1991 to 2005, were unique to the dataset-A. Area 8 in October, November and December from 1993 to 2005 were unique to the dataset-B.

Nominal CPUEs by Area are shown in Fig. A1. Large differences in the nominal CPUE between the dataset-A and B (CPUE_A < CPUE_B) were observed in Area 4 from 1993 to 1994 and in Area 9 from 1993 to 1994. Opposite difference (CPUE_A > CPUE_B) was observed in Area 8 from 1998 to 2000.

The nominal CPUE by Area, year and month for the dataset-A were very low in April, July and August of Area 4, and in April, August and September of Area9 (Fig. A2). These Area/month were outside of the fishing season for SBT and had effect to the nominal CPUE of all Areas in the dataset-A much lower than that of the dataset-B. Because the nominal CPUE in Area 7 in July was as high as that in the SBT fishing season, there were little difference in the nominal CPUE by both the datasets.

The nominal CPUE by Area, year and month for the dataset-B were slightly lower during October to December than in September in Area 8 (Fig. A3). The Area/month had effect to the nominal CPUE of all Areas in the

dataset-B slightly lower.

Therefore, it can be point out the basic differences between the two datasets. The dataset-A included a number of longline operations NOT for SBT, as a consequence provides lower nominal CPUE of SBT than in the dataset-B. The dataset-B included later half period of the Area 8 which is the one of the major fishing ground, and the dataset-B consisted mainly of longline operations for SBT.

By the way, similar results can be obtained using the 5x5, month data. Results by the 5x5, month data are attached in the Appendix.

Several CPUE standardizations

In Figure B0, the following ANOVA model (i.e. explanatory variables) in Equation (1) was used in both datasets (A and B). Two CPUE trends seem to be different in Figure B0 (See Figure 12 of p.30, Report of the second CPUE modeling workshop).

$$\log(\text{CPUE}+0.1)=\text{intercept}+\text{year}+\text{area}+\text{month}+\text{VesselID}+\text{HPB}+\text{observer}+\text{year}^*\text{observer}+\text{error}, \text{error} \sim N(0, \sigma^2) \quad (1)$$

However, in Figure B0, selected “Core-Vessels” were only used and the LSMENAS (least square means) of the year*observer without observer (i.e. in the case that scientific observers are not on board) were extracted as the estimated CPUE year trend. (Remark) This is an apparent mistake statistically and LSMEANS of the year effect should be extracted.)

Thus, since the starting point of discussion was wrong, we modified this point and used all vessels because which include more information. We also extracted the LSMEANS of year effect as the standardized year trend of CPUE for SBT using same Equation (1) in Figure B1.

CPUE year trends in Figure B1 are rather different from those in Figure B0 and the CPUE trends in two datasets in Figure B1 seem to be still different.

Next, we computed the standardized CPUE year trends using Eqn.(2), in which the main effect of observer and observer-related interactions (year*observer) are deleted from Equation (1). Formula (2) becomes a simple model using only main effects.

$$\log(\text{CPUE}+0.1)=\text{intercept}+\text{year}+\text{area}+\text{month}+\text{Vessel-ID}+\text{HPB}+\text{error} \quad (2)$$

In Figure B2, the year trends of standardized CPUE in both datasets are still different especially 1993-1994. Therefore, we check the CPUE trends

deleting the data for 1992-1995 in Formula (2) (See Figure B3). As a result, we obtained the similar trends from 1996 to 2005. Figure B4 shows the year trends of nominal CPUE in two datasets, where the gap of CPUE for 1993-1994 is seen as well as in Figure B2 and the year trends of nominal and standardized CPUE in the dataset-A (shown in Figure B2 and B4) is similar and those in the dataset-B is quite different.

At last, we applied more complicated model by Equation (3) with some interactions including the random effect because it seems not to be performed the corrections by CPUE standardization through the simple model using only main effects.

$\log(\text{CPUE}+0.1)=\text{intercept}+\text{year}+\text{area}+\text{month}+\text{Vessel-ID}+\text{HPB}+\text{observer}+(\text{year}*\text{observer})+(\text{year}*\text{area})+(\text{year}*\text{area}*\text{month})+\text{error}$, $\text{error} \sim N(0, \sigma^2)$ (3)
where $\text{year}*\text{area}*\text{month}$ is a random effect and other factors are fixed effect.

Estimated CPUE year trends obtained from the Equation (3) in two both datasets are shown in Figure B5 and those two trends seem to be different. The reason why the range of the confidence interval is wider than that in other figures is considered that the random effect is included into the model.

Reliability check of two datasets by the validation

We checked the reliability of both datasets (Dataset-A and B), which datasets has better performance from the statistical viewpoint, based on the simple validation. The procedure of the calculation applied for each datasets (A and B) is as follow:

1. Divided the all records (in both datasets) into two sub-datasets randomly, 80 percent of training data and 20 percent of the data for verification.
(Remark) We regarded the latter sub-dataset as missing data in this step)
2. After estimating unknown parameter by Equation (ANOVA model) only using the training data set, we computed the goodness of fit using the sub dataset for verification, which shows the difference between observed CPUEs and the corresponding predicted (i.e. obtained from the Equation (1)) ones, based on the mean absolute error (MAE) and Pearson's correlation coefficient.

Table B1 shows the values of mean absolute error and Pearson's correlation coefficient between observed and the corresponding estimated CPUE in the part of data for verification in two datasets. In addition, the plots of observed and the corresponding predicted CPUE are shows in Figure

B6. Judging from these values, the dataset-B is more robust/stable than dataset-A statistically.

Discussion

Temporary conclusions obtained from nominal and standardized CPUE are as follows:

- Differences between the two datasets in terms of the spatio-temporal coverage and nominal CPUE were observed in Area4 (April, July and August), Area 9 (April, August and September) and Area 8 (October-December).
- Both the datasets would have different merits and demerits in terms of reflecting the state of the stock abundance to CPUE. Using data only operations targeting for SBT seems to be a concern. Including a number of operations where and when few SBT were caught such as northern half of Area 4 also seems to be a concern. It should be investigated more comprehensively and in detail what kind of spatio-temporal range to be chosen is appropriate.
- Results from the shot-by-shot data were similar to that from the 5x5, month data. Further investigation based on 5x5, month data is possible and seems appropriate at least to some extents.
- The gap of year trends between the dataset A and B in the early 1990s especially for 1993-1994 seems to be still large in the standardized CPUE.
- Extracted year trends of standardized CPUE is dependent upon the model (i.e. explanatory factors included into the ANOVA model) utilized.
- The difference of the year trend between nominal and standardized CPUE seems to be rather similar in the dataset-A (area4-9 and month4-9) and quite different in the dataset-B (Japanese fishing season and zone).
- As a result of model validation, the dataset-B is more robust and stable than dataset-A statistically.

Acknowledgement

We acknowledge Prof. John Pope, Dr. Jim Ianelli, Mr. Naozumi Miyabe and Mr. Shigeyuki Kawahara for their useful comments.

References

Anonymous. 2007. Report of the Second CPUE Modelling Workshop. 21-25

- May 2007. Shimizu, Japan. 43pp.
- O. Sakai, T. Itoh and Y. Narisawa. 2007. Review of Japanese SBT fisheries in 2006. CCSBT-ESC/0709/ SBT-Fisheries/Japan. 46pp.

—Tables and Figures—

Table A0 Difference of the characteristic between datasets (A & B) and LL1

	Two datasets (A and B)	LL1 (used in the CCSBT)
Data resolution	Shot-by-shot	Aggregated by 5x5/month
Age configuration	All included	4plus (4+)
Joint venture vessels*	Not included	Included

*Data from the vessels of Joint venture between Japan and Australia or NZ.

Table A1 Spatio-temporal coverage of the dataset A and B made by shot-by-shot data

Number of hooks in thousands

DatasetA										DatasetB										$d=(a-b)/a$		$e=(a-c)/a$				
Area	Year	Month								a	Month								b	c	sum (Apr-Sep)	Excess A	Excess A (Apr-Sep)			
		4	5	6	7	8	9	Sum	4		5	6	7	8	9	10	11	12								
4	1991	438	1172	1422	1802	840	151	5824			703	1422	1802							3927	3927		33%	33%		
	1992	580	1428	1978	2516	763	194	7458			919	1978	2516							5413	5413		27%	27%		
1993	2150	2176	2522	2397	1238	86	10568			978	2522								3499	3499		67%	67%			
1994	1640	3492	2660	1263	942	106	10103			1778								1778	1778		82%	82%				
1995	1747	2292	2758	2138	1050	343	10328			1756	2077							3833	3833		63%	63%				
1996	1822	3960	4031	2076	1173	350	13413			2634	3446							6079	6079		55%	55%				
1997	2769	3745	3372	2088	708	82	12763			1261	3745	3372	782					9159	9159		28%	28%				
1998	1070	1752	2423	3679	987	38	11757			308	1752	4230	3679					9969	9969		15%	15%				
1999	666	693	1495	2803	648		6305			393	693	1495	2803	478				5862	5862		7%	7%				
2000	562	1358	2057	1979	262	66	6283			145	1358	2057	1979	16				5555	5555		12%	12%				
2001	421	339	1627	1932	493	100	4912			47	339	1627	1443					3456	3456		30%	30%				
2002	311	238	2751	3056	413	13	6780			54	238	2751	2376					5418	5418		20%	20%				
2003	305	949	2888	3368	657	295	8462			26	949	2888	3343					7205	7205		15%	15%				
2004	424	1108	2972	2959	1079	229	8771			64	1108	2972	2959					7102	7102		19%	19%				
2005	31	2073	3188	2033	3		7328			31	2073	3188	2033					7325	7325		0%	0%				
5	1991	114	62	1458	2872	529	231	5266															100%	100%		
	1992	242	1936	1429	182	48	3837																100%	100%		
1993	9	220	277	3			510																100%	100%		
1994		59					59																100%	100%		
1995	53	6	36	39	13		95																100%	100%		
1996		53	6	4	137		200																100%	100%		
1997			35	76			111																100%	100%		
1998			52	147	76		274																100%	100%		
1999				27	560	168	756																100%	100%		
2000					92	124	216																100%	100%		
2001			10	148	204	143	504																100%	100%		
2002			6	9	6	6	28																100%	100%		
2003		4		40	154	140	337																100%	100%		
2004				9	267	24	300																100%	100%		
2005					20																			100%	100%	
6	1991	2511	2795	537	37			5881																100%	100%	
	1992	1232	1535	198	288			3253																100%	100%	
1993	840	726	162	66			1794																	100%	100%	
1994	58	165	53				276																	100%	100%	
1995	117	267	256	59			699																	100%	100%	
1996																									100%	100%
1997	30	89	45				164																	100%	100%	
1998	9	223	159				392																	100%	100%	
1999	99	173	159	60			491																	100%	100%	
2000	18	18		25			61																	100%	100%	
2001	107	90	3				200																	100%	100%	
2002	25	91	54	3			173																	100%	100%	
2003	52	105	88				245																	100%	100%	
2004	1361	1631	6				2997																	100%	100%	
2005	1738	1127	10				2874																	100%	100%	
7	1991	1563	2589	1078	14	80	5324																	2%	2%	
	1992	808	1303	649	23	32	2815																	2%	2%	
1993	1224	1087	126				2438																	5%	5%	
1994		1909	983	3			2895																	65%	65%	
1995	86	988	1343	280			2697																	26%	26%	
1996		950	1102	358	197	86	2694																	30%	30%	
1997	569	3117	1749	164			5599																	0%	0%	
1998	1331	3685	1180	33	6	6235																		0%	0%	
1999	2209	4789	2951	271			364	10584															3%	3%		
2000	1943	2811	1699	37	63	6553																	1%	1%		
2001	2693	5081	3085	429			195	11482															2%	2%		
2002	2298	4608	1538	101			8545																	0%	0%	
2003	1918	3120	764		140	5941																	0%	0%		
2004	1361	312	6			2997																		0%	0%	
2005	1887	38				2874																		0%	0%	
8	1991	15	91	2440	3747	6293																	2%	2%		
	1992	66	1639	2730	4435																			-10%	1%	
1993		625	1158	1783																			77%	77%		
1994	18	67	833	3780	4699																		11%	20%		
1995	33	245	726	4387	5390																		-56%	19%		
1996	7	19	3	4549	4578																		-215%	1%		
1997			4370	4370																			-228%	0%		
1998	14	2251	4919	3509	10694																		-17%	67%		
1999	10	2457	4630	3328	10425																		26%	68%		
2000			434																							

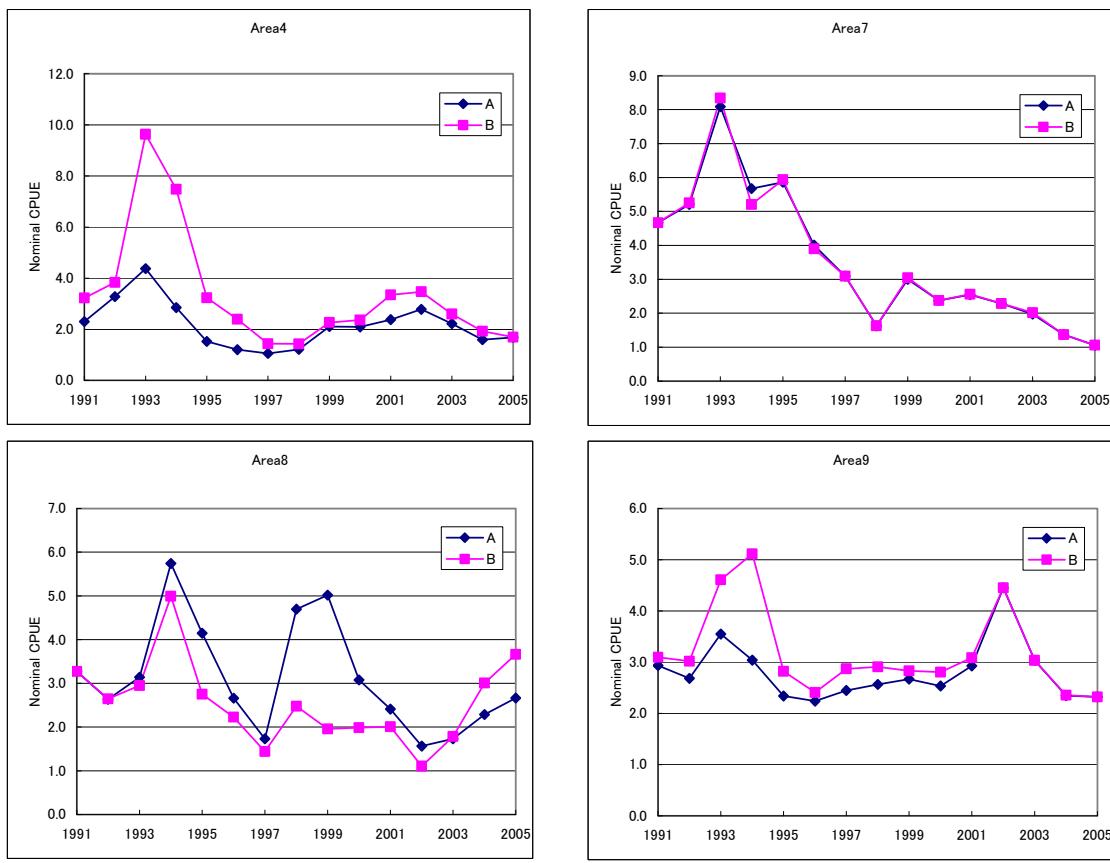


Fig. A1 Nominal CPUE by Area with the dataset A and B made by shot-by-shot data

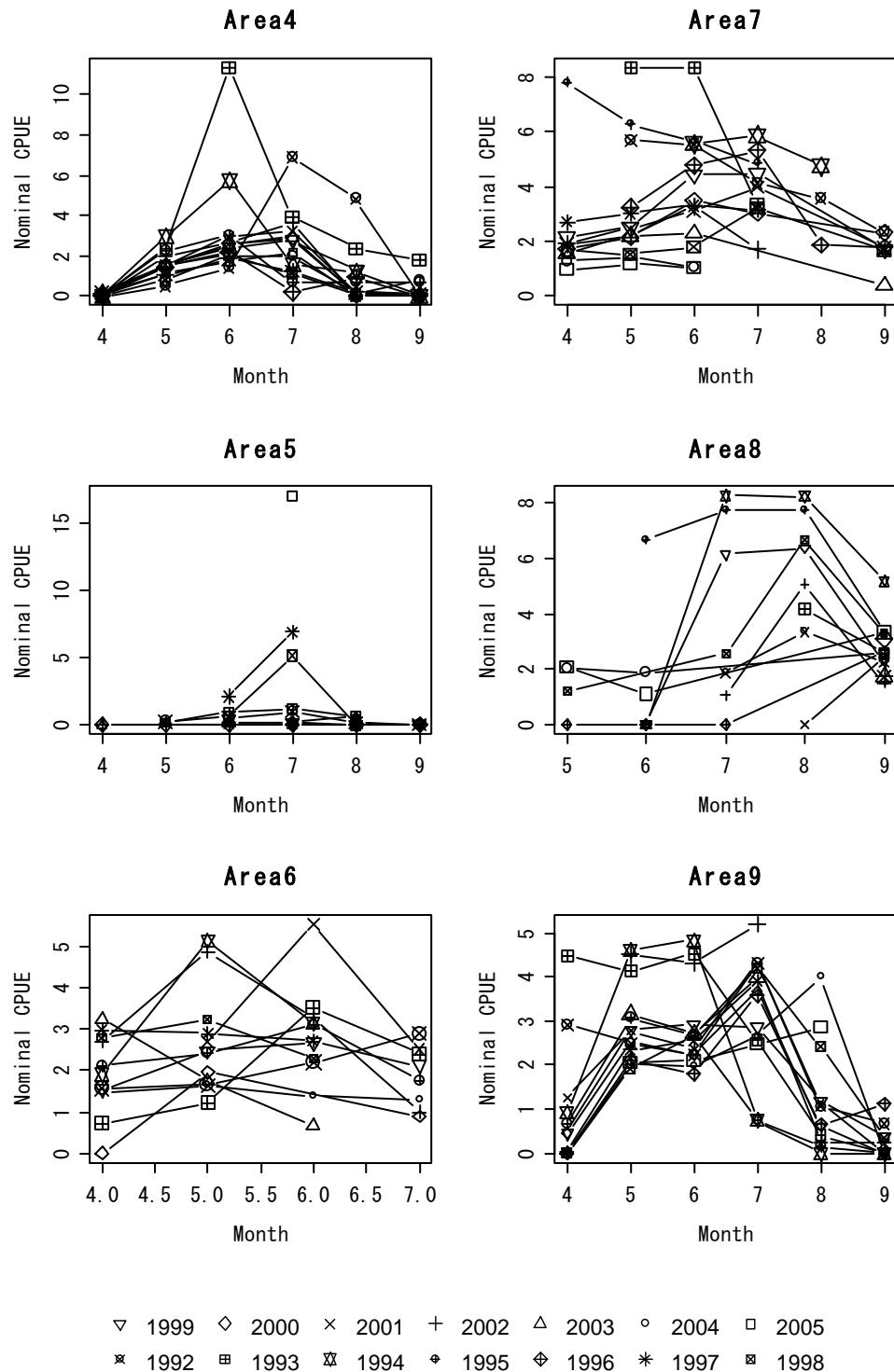


Fig. A2 Nominal CPUE by year, month and Area in the dataset A made by shot-by-shot data

Sizes of plots are proportional to the ratio of the effort in an area, year and month to the total efforts of the area.

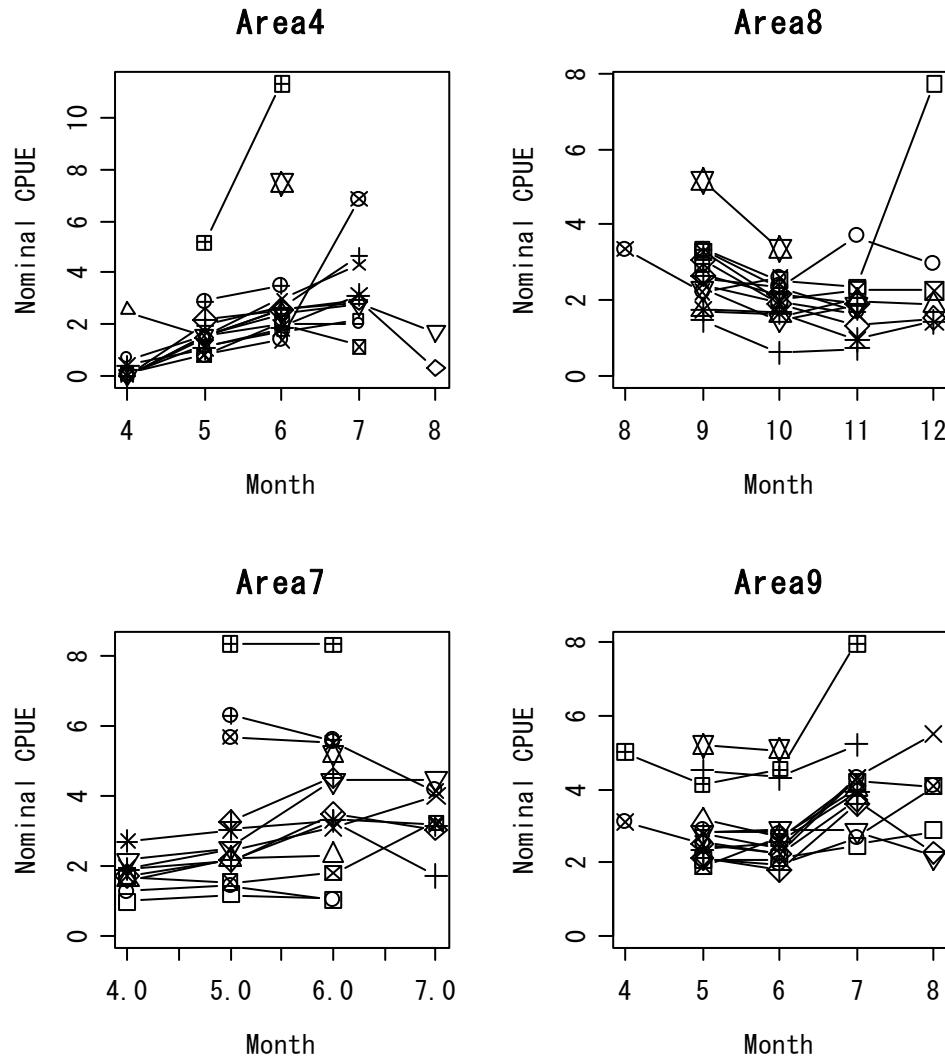


Fig. A3 Nominal CPUE by year, month and Area in the dataset B made by shot-byshot data

Sizes of plots are proportional to the ratio of the effort in an area, year and month to the total efforts of the area. Refer to the legend in Fig. 2A.

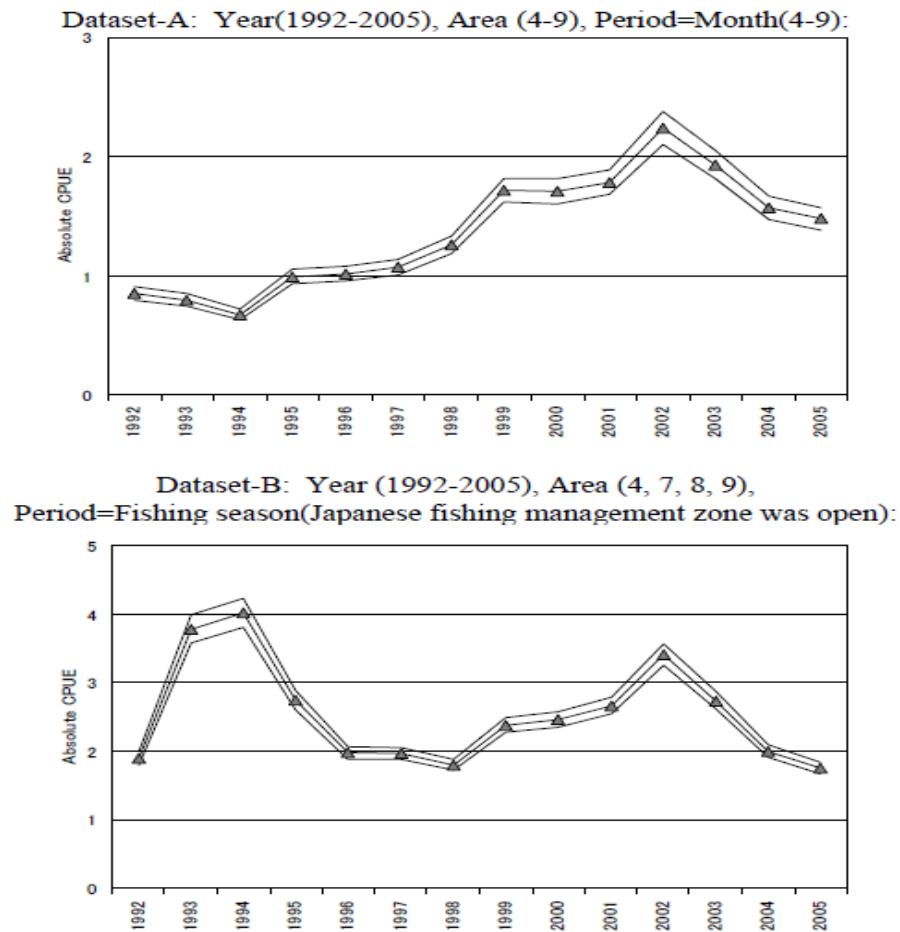


Figure 12. Comparison of model runs with core-vessels and different definitions of areas and times (Datasets A and B).

Figure B0 Estimated CPUE year trend using the LSMEANS of Year*Observer without observer and Core-vessels in the two datasets (A&B).

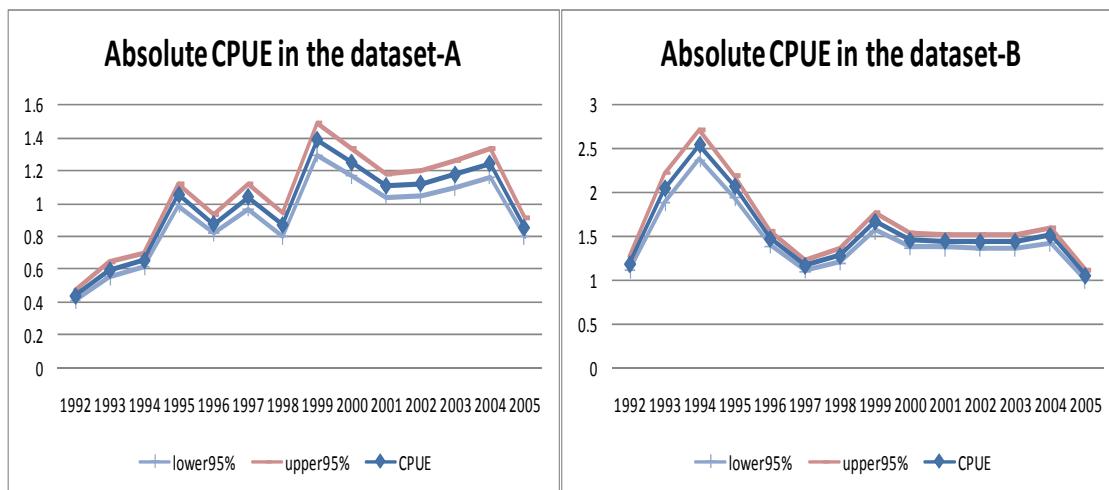
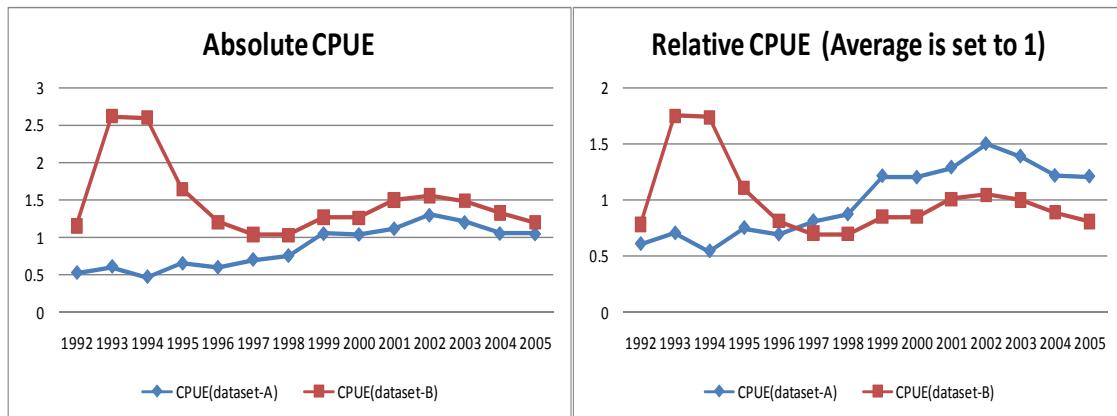
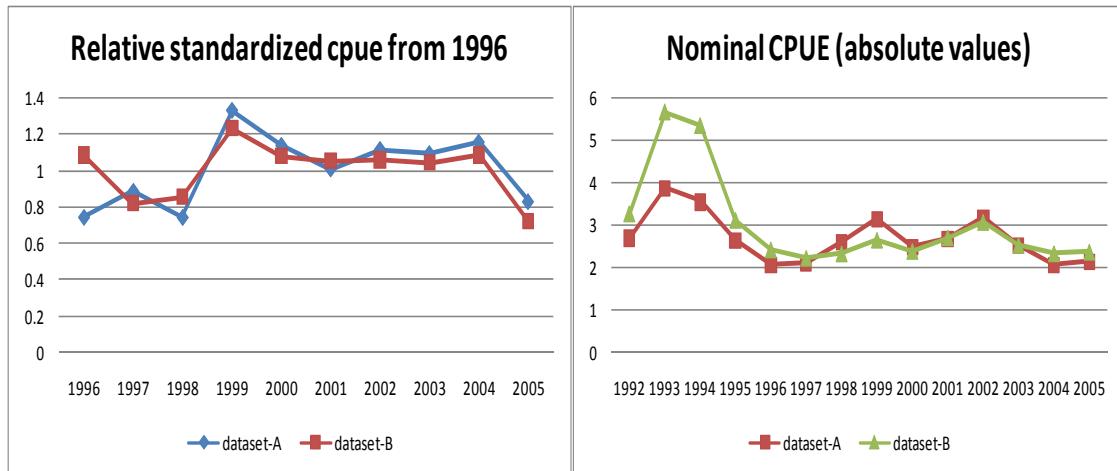
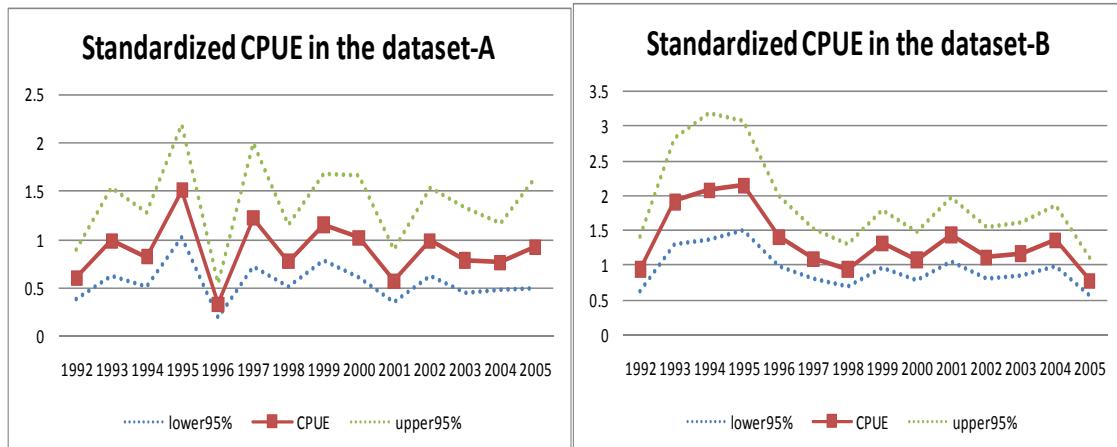


Figure B1 Estimated CPUE trend using LSMEANS of “Year” and all vessels.

Table B1 Values of mean absolute error and Pearson's correlation coefficient.

Dataset	MAE (Smaller is better)	Correlation (Larger is better)
Dataset-A	1.861	0.241
Dataset-B	1.618	0.349

**Figure B2** CPUE year trends in both datasets obtained from the Eqn.(2).**Figure B3** CPUE from 1996 in Eqn.(2). **Figure B4** Year trend of nominal cpue**Figure B5** CPUE year trends in both datasets obtained from the Eqn.(3).

Appendix Results made by 5x5 month data

Table A1 Spatio-temporal coverage of the dataset A and B by the 5x5 month data

Number of hooks in thousands

DatasetA										DatasetB												
Area	Year	Month								a sum	Month								b sum	c sum (Apr– Sep)	$d=(a-b)/a$	$e=(a-c)/a$
		4	5	6	7	8	9	10	11		4	5	6	7	8	9	10	11				
4	1991	440	1273	1485	1920	840	180	6137											4678	4678	24%	24%
1992	634	1517	2095	2774	1087	312	8418												6386	6386	24%	24%
1993	2499	2176	2531	2714	1580	193	11692												4707	4707	60%	60%
1994	1999	3997	2754	1277	1252	380	11659												2754	2754	76%	76%
1995	1662	2476	2997	1943	939	343	10359												5473	5473	47%	47%
1996	1822	4181	4417	2076	1173	350	14021												8599	8599	39%	39%
1997	2702	3884	3507	2076	667	82	12917												12168	12168	6%	6%
1998	1011	1888	4665	3891	896	3	12353											11454	11454	7%	7%	
1999	594	735	1663	2842	650	6483													6483	6483	0%	0%
2000	412	1284	1816	1833	26	5370													5370	5370	0%	0%
2001	319	337	1662	1855	332	36	4541											4172	4172	8%	8%	
2002	269	249	3031	3153	651	14	7365											6701	6701	9%	9%	
2003	259	976	2987	3368	570	264	8425											7590	7590	10%	10%	
2004	378	1246	3303	3155	1015	196	9294											8083	8083	13%	13%	
2005	84	2075	3192	2046	245	7642												7397	7397	3%	3%	
7	1991	1586	2612	1101	14	86	5399											5299	5299	2%	2%	
1992	800	1373	714	32	35	2954												2887	2887	2%	2%	
1993	1221	1060	201				2482											2281	2281	8%	8%	
1994		1552	648				2200											1552	1552	29%	29%	
1995	86	1003	1350	280			2719											2353	2353	13%	13%	
1996	989	1157	408	261	114	2929												2146	2146	27%	27%	
1997	600	3337	1862	167			5967											5967	5967	0%	0%	
1998	1566	4313	1410	33			6	7328										7322	7322	0%	0%	
1999	2363	5108	3200	271			365	11307										10943	10943	3%	3%	
2000	2032	3062	1916	49			69	7127										7058	7058	1%	1%	
2001	2737	5179	3132	444			195	11687										11491	11491	2%	2%	
2002	2603	5172	1744	134				9653										9653	9653	0%	0%	
2003	1955	3213	778				143	6089										5947	5947	2%	2%	
2004	1512	1814	6				3332											3332	3332	0%	0%	
2005	1744	1127	10				2880											2880	2880	0%	0%	
8	1991	15	95	2575	4102	6788					2575	4102							6677	6677	2%	2%
1992		68	1939	3010	5017						1939	3010	1304						6253	4949	-25%	1%
1993			543	1129	1672						1129		1129						33%	33%	33%	33%
1994		18	92	958	3760	4828					3760	531		4291		3760			11%	22%	22%	22%
1995	33	267	809	479	5887						4779	3115	1906	9800		4779			-66%	19%	-66%	19%
1996	7	19	3	4956	4984						4956	4769	5741	15465		4956			-210%	1%	-210%	1%
1997				4610	4610						4610	3831	4677	1946	15064	4610			-227%	0%	-227%	0%
1998	20	2555	5584	3985	12143						3985	4402	5031	879	14296	3985			-18%	67%	-18%	67%
1999	3	7	2781	5312	3449	11552					3449	2097	2398	7944	3449				31%	70%	31%	70%
2000			59	4610	4669						4610	4236	4868	2917	16631	4610			-256%	1%	-256%	1%
2001			38	4150	4188						4150	4266	4019	12436	4150				-197%	1%	-197%	1%
2002		3	138	4341	4482						4341	2364	965	7670	4341				-71%	3%	-71%	3%
2003				2801	2801						2801	2413	2360	917	8492	2801			-203%	0%	-203%	0%
2004		1176	312	1402	2890						1402	1675	2624	1851	7552	1402			-161%	51%	-161%	51%
2005		1887	38	1610	3534						1610	1993	2363	1420	7386	1610			-108%	54%	-108%	54%
9	1991	4119	6624	6882	6578	1014	519	25735										24202	24202	6%	6%	
1992	3673	5961	7551	6761	3058	1140	28144											23946	23946	15%	15%	
1993	3651	6990	8372	5383	2613	601	27611										24397	24397	12%	12%		
1994	1099	4511	6762	4166	1858	1179	19576										11273	11273	42%	42%		
1995	1417	8402	7148	1626	888	396	19877										15550	15550	22%	22%		
1996	581	8346	7245	5862	914	637	23585										21453	21453	9%	9%		
1997	598	803	7417	6343	1882	1338	25609										21790	21790	15%	15%		
1998	440	7239	7374	5768	2299	1344	24460										22677	22677	7%	7%		
1999	201	7176	7085	4864	1314	262	20896										20434	20434	2%	2%		
2000	33	5649	5405	5126	1185	302	17699										17364	17364	2%	2%		
2001	25	7277	7554	5974	875	299	22003										21679	21679	1%	1%		
2002		7032	7177	994			15202										15202	15202	0%	0%		
2003		7794	8211	1806			17811										17811	17811	0%	0%		
2004	48	6888	8445	8418	1334	3	25136										25085	25085	0%	0%		
2005		5495	7275	7737	3516	24023											24023	24023	0%	0%		

< -30% >60%
>20%

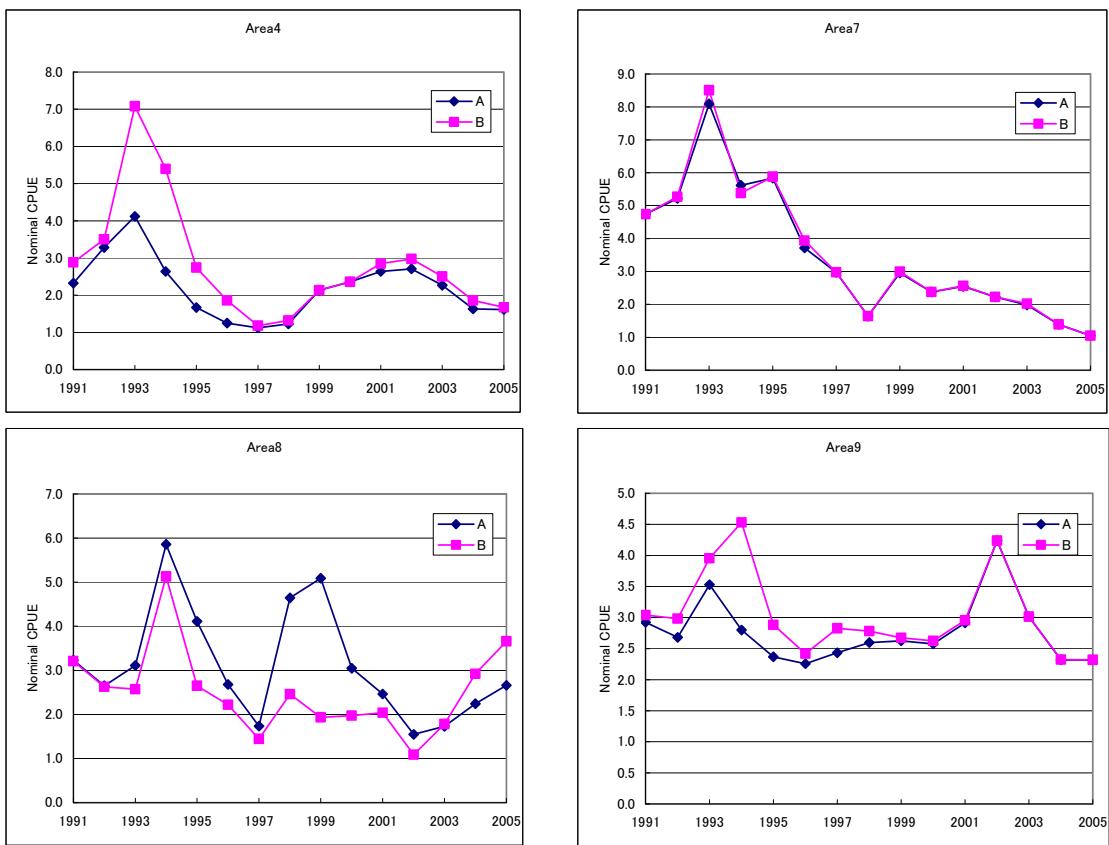


Fig. A1 Nominal CPUE by Area with the dataset A and B by the 5x5 month data

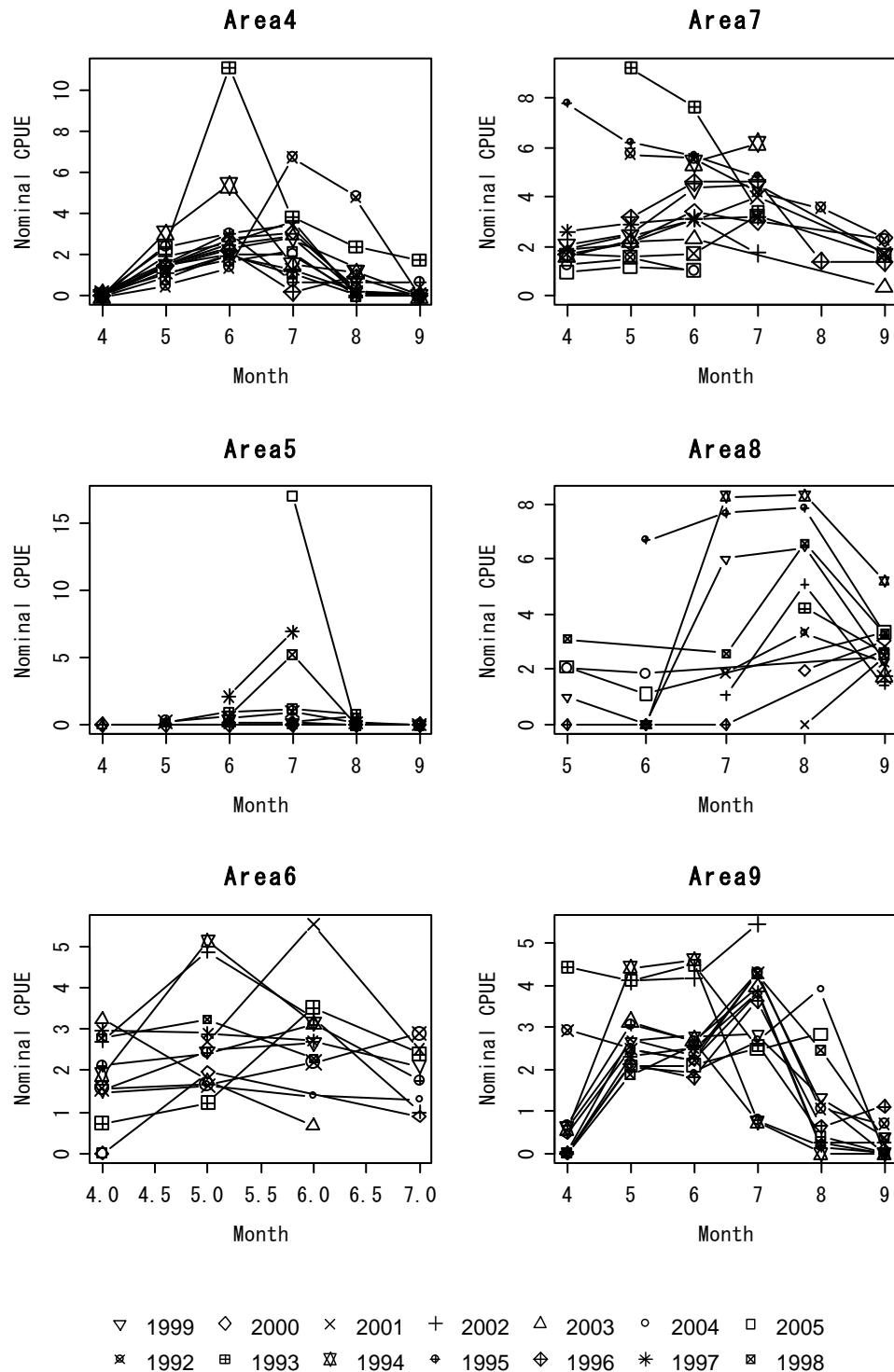


Fig. A2 Nominal CPUE by Area, year and month with the dataset A by the 5x5 month data

Sizes of plots are proportional to the ratio of the effort in an area, year and month to the total efforts of the area.

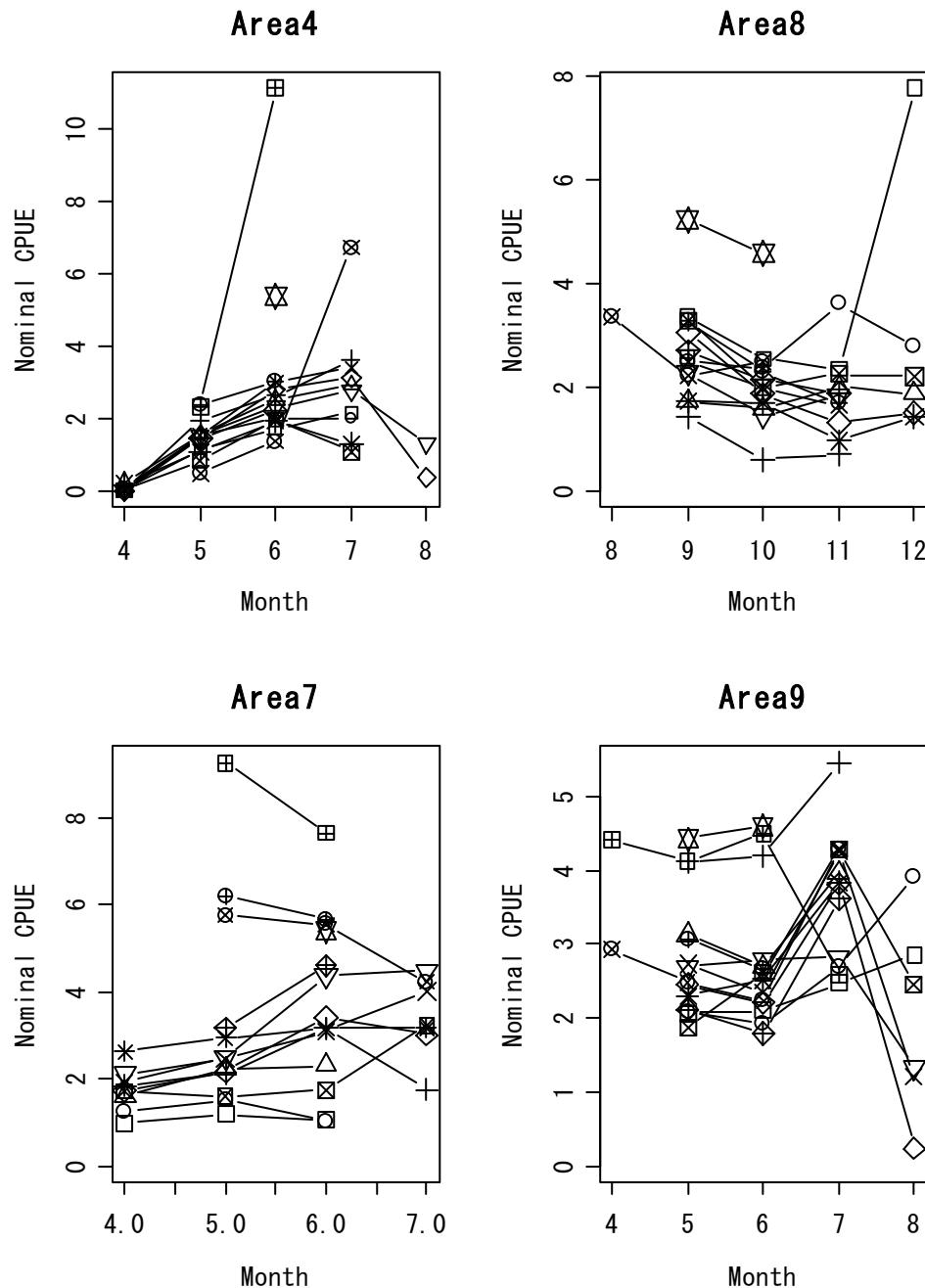
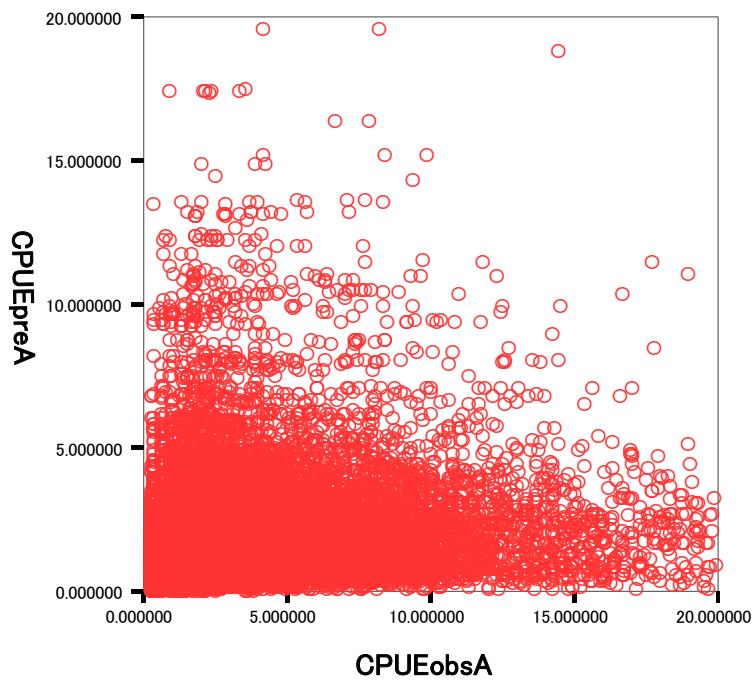


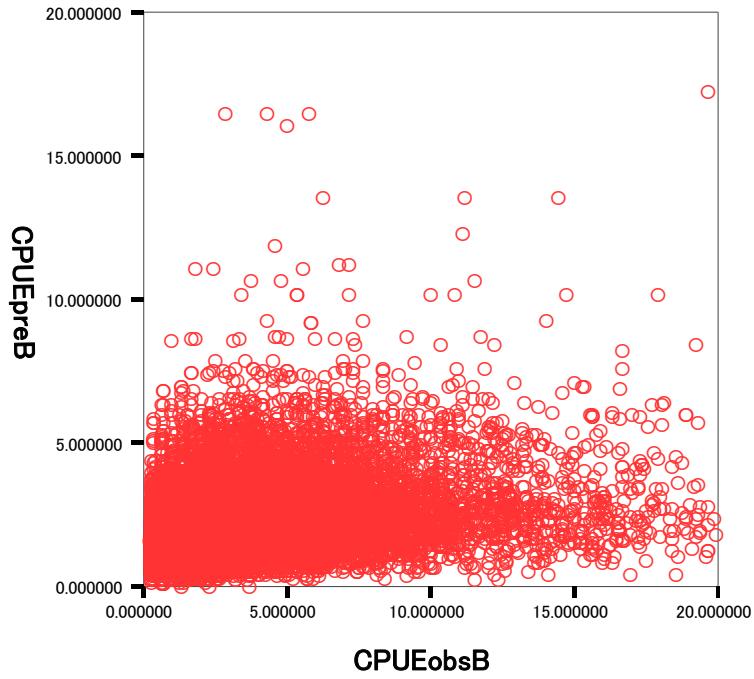
Fig. A3 Nominal CPUE by Area, year and month with the dataset B by the 5x5 month data

Sizes of plots are proportional to the ratio of the effort in an area, year and month to the total efforts of the area. See the legend in Fig. A3.



Dataset-A (Upper)

Dataset-B (Lower)

**Figure B6** Plot of observed and corresponding predicted CPUE in two dataset