Biogeochemistry **40:** 189–201, 1998.
© 1998 *Kluwer Academic Publishers. Printed in the Netherlands.*

The use of polynomial regression analysis with indicator variables for interpretation of mercury in fish data

GILLES TREMBLAY[1], PIERRE LEGENDRE[2], JEAN-FRANÇOIS DOYON[1], RICHARD VERDON[3] & ROGER SCHETAGNE[3]

[1]*Groupe Génivar, 5355, des Gradins, Québec (Québec) G2J 1C8, Canada. E-mail: gtremblay@genivar.com;* [2]*Département de sciences biologiques, Université de Montréal, C.P. 6128, succ. Centre-ville, Montréal (Québec) H3C 3J7, Canada;* [3]*Direction Environnement, Hydro-Québec, 75 boul. René-Lévesque ouest, 16ième étage, Montréal (Québec) H2Z 1A4, Canada*

Key words: binary variables, fish, impoundment, indicator variables, James Bay, mercury, polynomial regression, Québec, reservoirs

Abstract. Mercury levels in fish in reservoirs and natural lakes have been monitored on a regular basis since 1978 at the La Grande hydroelectric complex located in the James Bay region of Québec, Canada. The main analytical tools historically used were analysis of covariance (ANCOVA), linear regression of the mercury-to-length relationship and Student-Newman-Keuls (SNK) multiple comparisons of mean mercury levels. Inadequacy of linear regression (mercury-to-length relationships are often curvilinear) and difficulties in comparing mean mercury levels when regressions differ lead us to use polynomial regression with indicator variables.

For comparisons between years, polynomial regression models relate mercury levels to length (L), length squared (L^2), binary (dummy) indicator variables (B_n), each representing a sampled year, and the products of each of these explanatory variables ($L \times B_1$, $L^2 \times B_1$, $L \times B_2$, etc.). Optimal transformations of the mercury levels (for normality and homogeneity) were found by the Box-Cox procedure. The models so obtained formed a partially nested series corresponding to four situations: (a) all years are well represented by a single polynomial model; (b) the year-models are of the same shape, but the means may differ; (c) the means are the same, but the year-models differ in shape; (d) both the means and shapes may differ among years. Since year-specific models came from the general one, rigorous statistical comparisons are possible between models.

Polynomial regression with indicator variables allows rigorous statistical comparisons of mercury-to-length relationships among years, even when the shape of the relationships differ. It is simple to obtain accurate estimates of mercury levels at standardized length, and multiple comparisons of these estimations are simple to perform. The method can also be applied to spatial analysis (comparison of sampling stations), or to the comparison of different biological forms of the same species (dwarf and normal lake whitefish).

1. Introduction

Since the 1970's, many studies have shown the relation between impoundment of reservoirs and rise of mercury concentrations in fish (Potter et al. 1975;

Kelly et al. 1975; Abernathy & Cumbie 1977; Bruce et al. 1979; Bodaly et al. 1984; Messier & Roy 1987; McMurtry et al. 1989; Brouard et al. 1990; Monteiro et al. 1991). Some authors have also demonstrated that a rise of mercury in fish in natural lakes may be related to the anthropogenic sources and activities (Håkanson et al. 1988; Lucotte et al. 1995). Most of the time, the relation between mercury concentration and length, weight or age of fish was described by linear regression and comparison between year or sampling station (if any) was made with covariance analysis. In some instance comparisons were based on more simple statisticals tools like analysis of variance and multiple range tests of means mercury levels.

As for other water bodies contaminated with mercury, a monitoring program for fish was implemented in the La Grande hydroelectric complex, located in the James Bay region of Québec (Canada). A first attempt was made in 1987 at formulating an appropriate statistical procedure for analysis of data (Brouard et al. 1987), which include length, weight, sex, age, year and sampling station. The goals were to follow the temporal evolution of mercury levels in fish, which are related to length, and to compare the levels and relationships of mercury to length within reservoirs, and between reservoirs and natural lakes. The analysis should also be able to provide information on the duration of elevated mercury levels in reservoir fish.

Some problems arose using earlier methods. Linear regression equations did not always fit the data well. In addition, the square root transformation of mercury concentration was not always appropriate. Analysis of covariance, using fish length as the covariate, was also difficult to apply because for many cases the conditions of equality of the variances and slopes among the linear regression models were not met. In these cases, the alternative SNK procedure (no regression technique involved in this test) was also unsatisfactory because the mercury levels being compared did not correspond to fish of similar lengths.

The revised method of data analysis includes (1) transformation for normality and homoscedasticity treated on a species-specific basis, (2) a non-linear generalisation of the analysis of covariance (Legendre & McArdle 1997), involving polynomial regression with indicator variables of mercury-to-length relationships; (3) comparison of confidence intervals of mercury levels estimated at standardized length by the polynomial regression; and (4), when mercury levels have returned to background concentrations, power analysis following Student's *t*-test between levels found in impoundments and in natural lakes of the same region.

This paper describes the former and new methods stressing the differences and the points of improvement. We also describe polynomial regression analysis with indicator variables and present a few examples of the application of

the method. This method, with a fully worked example, is described in a guidance document (Tremblay et al. 1996) which is available both in French and English by request from the authors.

2. Description and comparison of the methods

2.1 *Former procedure (1987)*

Figure 1 shows the flow diagram of data analysis under the former statistical approach (Brouard et al. 1987). Three decision points are found in the diagram, based on five conditions of application for analysis of covariance. The first decision point considers normality and homoscedasticity, which are required for parametric methods such as analysis of covariance and SNK comparisons. The second decision point tests the homogeneity of residual variances among regression lines. If this condition was not met, non-parametric Kruskall-Wallis tests were used to compare mean mercury levels.

The third decision point occurs when testing for equality of slopes of the regression lines. If that condition is met, analysis of covariance can be used to compare mercury levels by testing for equality of the intercepts (elevation). If slopes are not equal, a situation often encountered, the analysis of mercury levels can only be made by the SNK multiple comparison test. Because of differences in mean total lengths, there were discrepancies in some cases between mean mercury levels compared by the SNK procedure and the estimated mercury levels at standardized length. An example is presented in Table 1. Mercury levels reported in column "ST" (estimated from linear regression) indicate that the statistical decision in column "Mean" is incorrect, the value from 1989 being in fact the lower. Interpretation of the temporal evolution of mercury as well as the spatial comparisons of mercury levels could thus be seriously biased.

Even when the analysis of covariance could be applied, estimation of mercury levels at standardized length was not realistic in many cases because of the lack of fit of a linear model to the data, particularly a few years after impoundment of the reservoirs, when the mercury-to-length relationships often became curvilinear (Figure 2).

2.2 *New procedure (1995)*

The new approach is much less restrictive in its conditions of application. There is only one decision box in the diagram (Figure 3), referring to assumptions of normality and homogeneity of the variances. The other conditions required by the analysis of covariance (equality of slopes and variance) are not

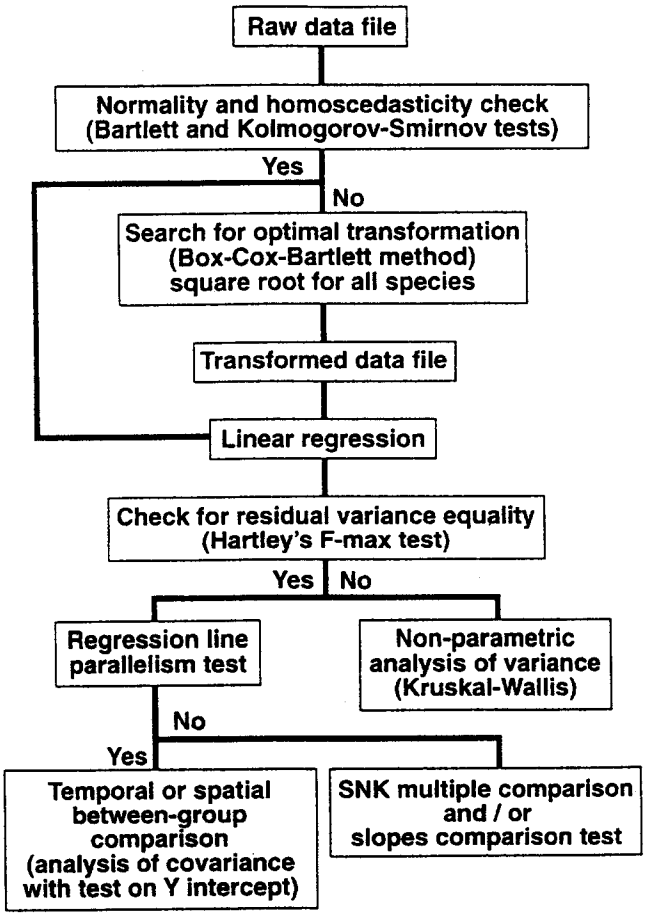


Figure 1. Data analysis flow diagram under the former statistical approach (1987).

Table 1. Temporal evolution of mercury levels in lake whitefish from the Caniapiscau river (Calcaire station, 1989 to 1993).

		Total mercury (mg/kg)					Total length (mm)		
Year	*N*	ST[1]	Mean[2]	Min.	Max.	Coeff. var. (%)	Mean	Min.	Max.
1989	11	0.12	0.29(a)[3]	0.08	0.54	46	479	370	555
1991	23	0.18	0.18(b)	0.05	0.45	69	400	179	578
1993	30	0.15	0.20(b)	0.07	0.37	48	443	318	537

[1]: Mercury levels estimated for standardized length by linear regression.
[2]: Arithmetic means levels tested with the SNK procedure.
[3]: Mean levels with different letters are significantly different at the 5% level.

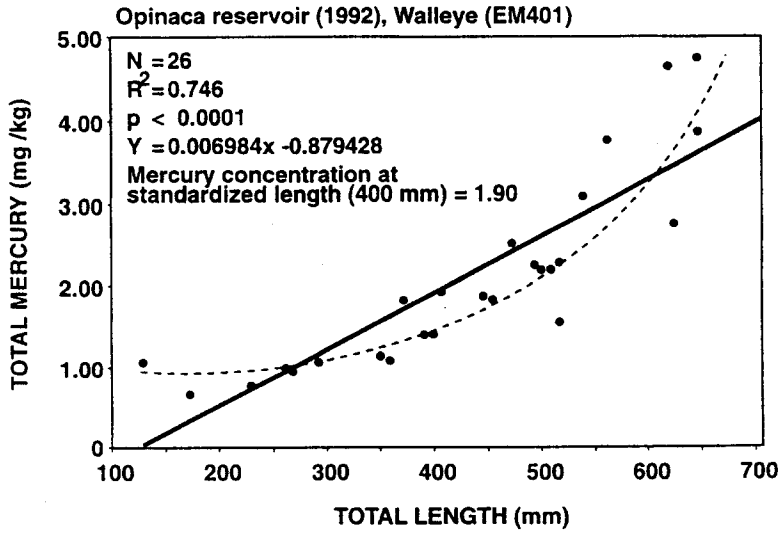


Figure 2. Inadequacy of the linear model (straight line) for the walleye (*Stizostedion vitreum*). A quadratic model (curve) fits this data better.

required for polynomial regression, so recourse to non parametric methods is not necessary. Estimated mercury levels, obtained by polynomial regressions for standardized lengths, are compared through their 95% confidence intervals.

In temporal analysis (between years), when mercury levels seem to have returned to normal conditions (i.e., before impoundment of a reservoir), a Student's *t*-test is run between the reference condition (natural mercury level) and the levels to be tested. If the test fails to show any difference, a minimal detectable difference is determined (based on observed natural variation) and a power analysis is run to determine the power offered by the test considering sample size (Cohen 1988). If the observed difference is less than the minimal difference determined above, and the power of the test is at least 0.80 (1-power is the probability of making a type II error by incorrectly accepting the null hypothesis of equality of means; Zar 1984), the level of mercury is considered to have returned to normal conditions.

3. Polynomial regression with indicator variables

3.1 *Construction of the model*

Any problem of analysis of variance or covariance can be recast into a multiple regression analysis (Draper & Smith 1981; Freund & Littel 1981; Snedecor &

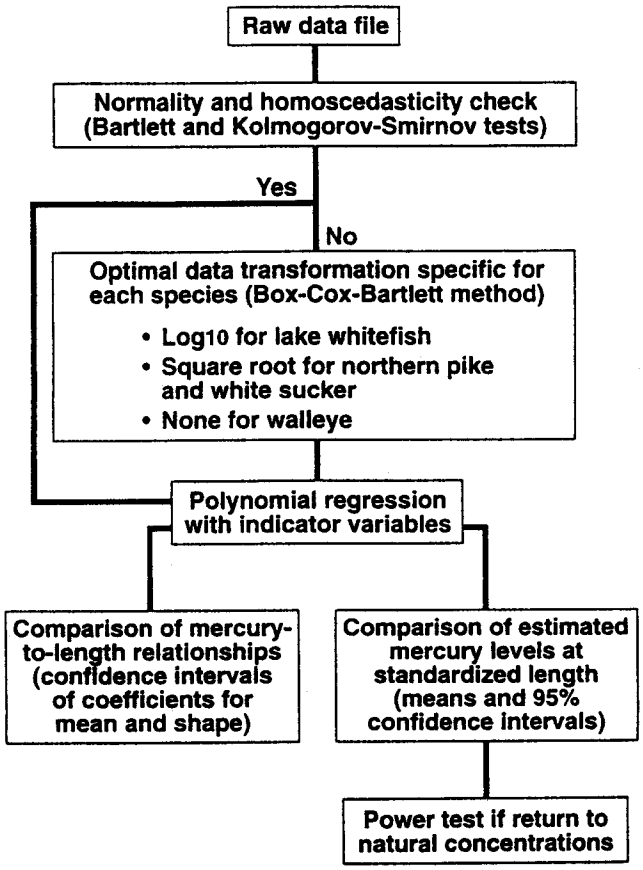


Figure 3. Data analysis flow diagram under the new statistical approach.

Cochran 1980). Polynomial regression analysis offers a way of generalizing the analysis of covariance to nonlinear situations, and this method is adequate for the requested analysis. The main advantage of the new approach is to explicitly model the mercury-to-length relationships being compared.

In polynomial regression analysis with indicator variables, the aim is to develop a general equation describing the relationship between mercury levels and length, for all years involved in the analysis, while taking in account not only the differences in mean mercury levels but also the different shapes of the mercury-to-length relationships. This goal is achieved by introducing a set of binary (dummy) variables into the model. Each variable describes a particular year or station. The following example is for a temporal analysis, but the technique can also be applied to spatial comparisons.

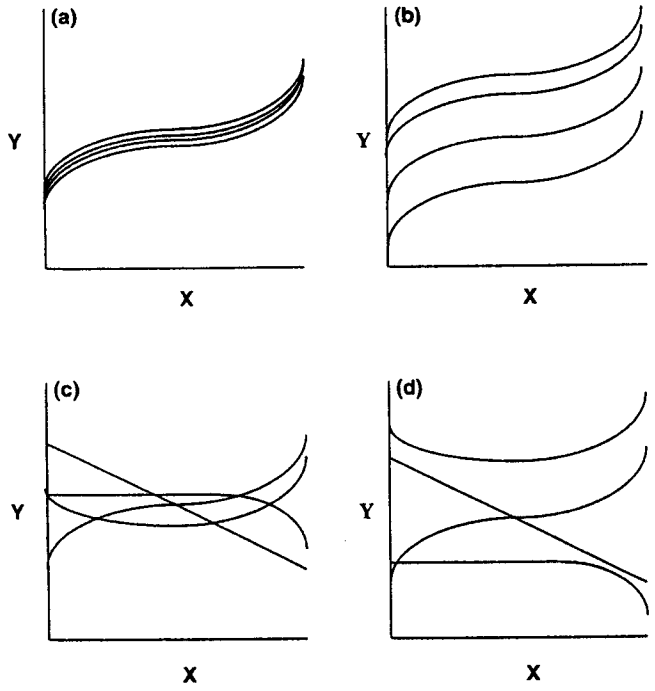


Figure 4. Polynomial regression with indicator variables makes it possible to distinguish the following four cases: (a) All mercury-to-length relationships can be described by a single equation (the curves should actually be drawn on top of one another; they are only separated to show that there are several curves). (b) Differences in means but no difference in shapes. (c) Differences in shapes but no difference in means. (d) Differences in means and shapes. The shapes of the curves in this figure are not intended to represent mercury-to-length relationships.

Mercury levels in fish can be adequately described by a polynomial function of their length. In this example, second degree polynomials only (quadratic functions) were used because we found them sufficient to describe the relationships; polynomes of higher order may be necessary for other types of data. For a single group of fish, or for different groups with the same mean and shape of the mercury-to-length relationship (Figure 4a), the quadratic model is:

$$\mathrm{Hg} = a + bL + cL^2 + \varepsilon \tag{1}$$

where Hg is the mercury level, L represents fish length, and a, b and c are the parameters (coefficients) of the model; ε is the usual error term of the regression model, describing the difference between the observed and predicted values.

Differences in mean mercury levels among years (Figure 4b) can be expressed by introducing binary (dummy) variables (B_1, B_2, ...) into the

model, with each binary variable representing a sampled year. The reference year, before impoundment, is omitted from the model because when all the other binary variables are set to 0, the only remaining choice is 1 for that binary variable. The following equation illustrates a model for two years only:

$$\text{Hg} = a + bL + cL^2 + dB + \varepsilon \tag{2}$$

When the dummy variable B is set to 0, the equation represents the first-year's mercury-to-length relationship, and when set to 1, the model describes the second-year's curve. The new parameter d estimates the difference in mean mercury levels between the two years. The reference year (usually the first) is given the value $B = 0$. This has a useful consequence: the first three terms of the equation model the mercury-to-length relationship of the reference year, while the fourth term expresses the fact that the second-year curve may have a different mean value, although it has the same shape as the first-year curve.

Differences in shape among years (Figure 4c) can be modelled by adding new terms to equation 1. These terms are obtained by multiplying the binary variables with the length variables (L, L^2). For simplicity, the following equation illustrates a model containing a single binary variable B (two years only):

$$\text{Hg} = \text{a} + \text{bL} + \text{cL}^2 + \text{eBL} + \text{fBL}^2 + \varepsilon \tag{3}$$

Using the same convention as above ($B = 0$ for the reference year), the first three terms of the equation model the shape of the mercury-to-length relationship for the reference year, while the next two terms allow the model to express a difference in shape for the second-year curve.

Combining equations 2 and 3 allows the modelling of differences in mean as well as in shape among years (Figure 4d). Considering two years only, the model is:

$$\text{Hg} = a + bL + cL^2 + dB + eBL + fBL^2 + \varepsilon \tag{4}$$

The same model, extended to several years, would include one binary variable fewer than the number of years being studied:

$$\begin{aligned}\text{Hg} = {} & a + bL + cL^2 + dB_1 + eB_2 + \ldots + oB_n + pB_1L + qB_1L^2 + \\ & rB_2L + sB_2L^2 + \ldots + yB_nL + zB_nL^2 + \varepsilon\end{aligned} \tag{5}$$

where $a \ldots z$ are parameters of the model. This equation allows modelling of all possible combinations of differences in mean and shape among several

years of data (Figure 4d). Using a step-by-step backward elimination procedure, terms that do not significantly contribute to the coefficient of determination (R^2, which is the proportion of variance explained by the model) are removed in order to ease the use of equation. Note that lengths are first centered by subtracting the mean, before calculation of the squared length term, in order to reduce the linear dependence between L and L^2 which creates a side effect on the stepwise procedure and on the regression (collinearity). The resulting general equation can be split into parts, as indicated above, to describe specific mercury-to-length relationships for the various years.

3.2 *Comparing the curves and the estimated mercury at standardized length*

The equations for the various years can be compared directly, since they have the same basic structure (quadratic in our models). They are compared on the basis of their means (coefficient of binary variables) and their shapes (coefficient of L and L^2 terms when present; equations with different terms are differents). A set of letters may be used to summarise, in a table, the possible groupings of years as to their mean mercury levels and shapes, based upon comparisons of the confidence intervals of the model parameters (Figure 5).

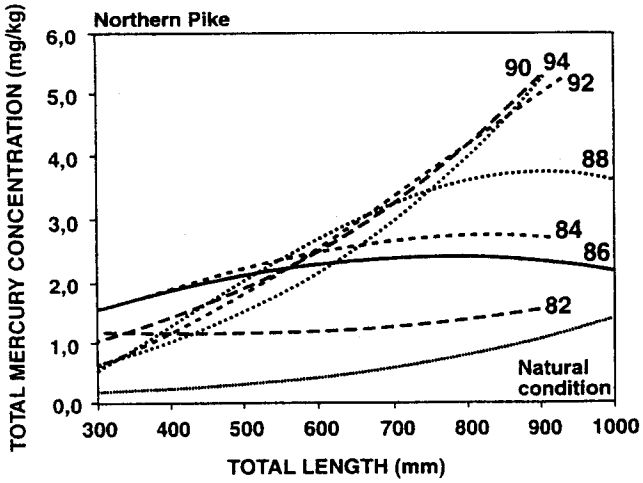
The mercury level at standardized length for each year can be estimated, along with the 95% confidence interval of its mean, from the resulting equation, using standard matrix calculations. The estimated mean mercury levels at standardized length are compared by their confidence intervals. Letters may be used to indicate, in a table, the possible grouping of years based upon confidence intervals comparisons as it is done for comparing the equation coefficients (Figure 6). In this example, results indicate that years 1984 and 1986 on the one hand, and 1988 through 1992 on the other, have undistinguishable mercury levels at the standardized length of 700 mm

While the examples given above concern temporal analysis, the method can also be used for spatial analysis or for comparing different biological forms of the same species.

3.3 *Normality and homoscedasticity*

Parametric statistical methods require that the conditions of normality and homogeneity of the variances of the dependant variable (mercury) be met before applying those techniques. Data transformations are often used to correct for possible deviations from these conditions, but in some cases it is not possible to find an adequate data transformation, or one finds that the optimal transformation is not the same among data sets.

Polynomial regression with indicator variables is moderately robust to deviations from these conditions. However, in order to come as close as



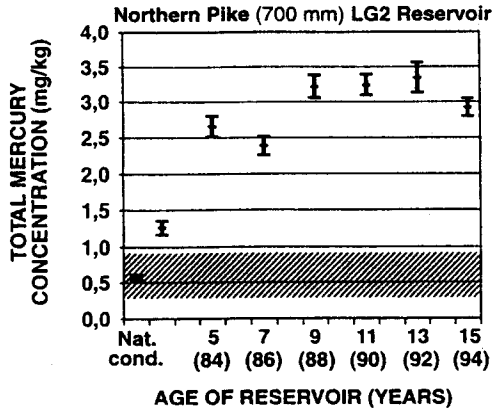
Age of res. (year)	Nat. cond.	3 (82)	5 (84)	7 (86)	9 (88)	11 (90)	13 (92)	15 (94)
N	373	147	148	150	139	114	85	167
Elevation	e	d	bc	c	ab	a	a	b
Shape	b	c	de	e	d	a	de	a

Note : Different letters indicate significantly different mean or shape (p<0,05)

Figure 5. Temporal evolution of mercury-to-length relationships for the northern pike (*Esox lucius*) from the La Grande 2 reservoir from 1982 to 1994 (noted 82–94).

possible to optimal analysis conditions, the three major species in the monitoring program (lake whitefish, *Coregonus clupeaformis*; northern pike, *Esox lucius*; and longnose sucker, *Catostomus catostomus*) have been analysed for optimal transformation using the Box-Cox-Bartlett procedure (Sokal & Rohlf 1995), using subsamples drawn from the data sets accumulated for reservoirs La Grande 2, La Grande 3 and Opinaca for the years 1986, 1988 and 1990. This technique selected the transformation that optimises both normality and homoscedasticity. The logarithmic (base 10) and the square root transformations appeared to be the best ones, when one is required. These two transformations, or no transformation at all, have been checked systematically for every species involved in the monitoring program; we selected in each case the one that appeared more regularly to be the best one for that species.

For example, the optimal transformation for lake whitefish and lake trout (*Salvelinus namaycush*) was the log (base 10); the square root was the best



Comparison of mean mercury levels estimated for standardized length at a 5% significance level.								
Age of res. (year)	Nat. cond. e	3 (82) d	5 (84) c	7 (86) c	9 (88) a	11 (90) a	13 (92) a	15 (94) b
Estimated level	0,58	1,27	2,66	2,39	3,21	3,24	3,34	2,92
Lower limit	0,54	1,17	2,52	2,27	3,05	3,10	3,13	2,80
Upper limit	0,61	1,36	2,81	2,52	3,38	3,39	3,57	3,04
N	373	147	148	150	139	114	85	167

Note : Years with different letters have confidence intervals (95%) which do not overlap. Letter "a" is attributed to the highest value.

Figure 6. Temporal evolution of mean mercury levels and 95% confidence intervals estimated at the standardized length for the northern pike (*Esox lucius*) from 1982 to 1994 (noted 82–94).

one for the northern pike, brook trout (*Salvelinus fontinalis*) and longnose sucker; while no transformation was required for the walleye (*Stizostedion vitreum*). After modelling by polynomial regression with indicator variables, each adjusted model was checked by looking at a plot of the residuals, in order to detect discrepancies with respect to the conditions of application. No deviation large enough to adversely affect the results was detected. All figures are «back» transformed for the mercury variable, so confidence intervals may be asymetric with respect to their means.

4. Conclusions

Polynomial regression with indicator variables is a powerful tool for interpreting mercury levels in fish. It is much less restrictive than previous methods.

The model can use higher-order polynomials but sufficient flexibility has been obtained using the quadratic form. It allows rigorous statistical analyses of mercury-to-length relationships among years, even when the shape of the relationships differ. It is simple to obtain accurate estimates of mercury levels at standardized length, and multiple comparisons of these estimations are simple to perform. The method can also be applied to spatial analysis (comparison of sampling stations), or to the comparison of different biological forms of the same species. Normality and homoscedasticity of the distributions are still required, although the method is robust to moderate deviations from these conditions, which can be satisfied by applying appropriate transformations of the data.

References

Abernathy AR & Cumbie PM (1977) Mercury accumulation by largemouth bass (*Micropterus salmoides*) in recently impounded reservoirs. Bull. Envir. Contam. Toxicol. 17: 595–602

Bodaly RA, Hecky RE & Fudge RJP (1984) Increase in fish mercury levels in lakes flooded by the Churchill River diversion, northern Manitoba. Can. J. Fish. Aquat. Sci. 41: 682–691

Brouard D, Demers C, Lalumière R, Schetagne R & Verdon R (1990) Rapport synthèse. Évolution des teneurs en mercure des poissons du complexe hydroélectrique La Grande, Québec (1978–1990). Rapport conjoint vice-présidence Environnement, Hydro-Québec et Groupe Environnement Shooner inc. 100 pp

Brouard D, Verdon R & Legendre P (1987) Réseau de surveillance écologique du complexe La Grande. Stratégie d'échantillonnage et processus d'analyse des données de mercure dans la chair des poissons. Service Études et Recherches écologiques, Direction Environnement, Hydro-Québec, 10 pp

Bruce WJ, Spencer KD & Arsenault E (1979) Mercury content data for Labrador fishes, 1977–1978. Fish. Mar. Serv. Data Rep. 142: 258

Cohen J (1988) Statistical Power Analysis for the Behavioral Sciences, 2nd edn. Lawrence Erlbaum Assoc., Publ., Hillsdale, New Jersey

Draper NR & Smith H (1981) Applied Regression Analysis, 2nd edn. John Wiley and Sons, Inc., New York

Freund RJ & Littel RC (1981) SAS for Linear Models. A Guide to Anova and GLM Procedures. SAS Series in Statistical Applications, SAS Institute, Cary, North Carolina

Håkanson L, Nilsson Å & Andersson T (1988) Mercury in fish in Swedish lakes. Environmental Pollution 49: 145–162

Kelly TM, Jones JD & Smith GR (1975) Historical changes in mercury contamination in Michigan walleyes (*Stizostedion vitreum vitreum*). J. Fish. Res. Board Can. 32: 1745–1754

Legendre P & McArdle BH (1997) Comparison of surfaces. Oceanologica Acta 20(1): in press.

Lucotte M, Mucci A, Hillaire-Marcel C, Pichet P & Grondin A (1995) Anthropogenic mercury enrichment in remote lakes of Northtern Québec (Canada). Water Air Soil Poll. 80: 467–476

McMurtry MJ, Wales DL, Scheider WA, Beggs GL & Dimond PE (1989) Relationship of mercury concentrations in lake trout (*Salvelinus namaycush*) and smallmouth bass (*Micropterus dolomieui*) to the physical and chemical characteristics of Ontario lakes. Can. J. Fish. Aquat. Sci. 46: 426–434

Messier D & Roy D (1987) Concentrations en mercure chez les poissons au complexe hydroélectrique de La Grande Rivière (Québec). Natur. Can. (Rev. Écol. Syst.). 114: 357–368

Monteiro LR, Isidro EJ & Lopes HD (1991) Mercury content in relation to sex, size, age and growth in two scorpionfish (*Helicolenus dactylopterus* and *Pontinus kuhlii*) from Azorean waters. Water Air Soil Poll. 56: 350–367

Potter L, Kidd D & Strandiford D (1975) Mercury levels in Lake Powell. Bioamplification of mercury in man-made desert reservoir. Envir. Sci. 40: 2251–2259

Snedecor GW & Cochran WG (1981) Statistical Methods. Iowa State University Press, Ames, Iowa

Sokal RR & Rohlf FJ (1995) Biometry, 3rd edn. W.H. Freeman and Co., New York

Tremblay G, Doyon JF & Schetagne R (1996) Réseau de suivi environnemental du complexe La Grande. Démarche méthodologique relative au suivi des teneurs en mercure des poissons. Rapport conjoint Direction principale Communications et Environnement d'Hydro-Québec et Groupe-conseil Génivar inc. 33 pp. et annexes

Zar JH (1984) Biostatistical Analysis. 2nd edn. Prentice-Hall Inc., New York