Stockholm
University

Master's thesis

Department of Statistics

Masteruppsats, Statistiska institutionen

Nr 2019:01

Dioxin and PCB content in *European whitefish* – a generalized linear mixed model

Hannes Waldetoft

Master's Thesis 30 ECTS, VT 2019

Supervisor: Hans Nyquist

Abstract

A generalized linear mixed model is used to examine how a selection of explanatory variables are related to dioxins and PCBs in European whitefish in Swedish waters. The long term goal is to perform the fishing in a manner that ensures low levels of dioxins and PCBs in fish going for sale. Data are supplied by the project "Dioxins in oily fish - threats and potential for development of small scale coastal and lake fishing" by IVL Swedish Environmental Research Institute in collaboration with SLU Aqua and the Swedish Veterinary Institute. The dependent variable used is assumed to be gamma distributed. It is a measure of toxicity of the dioxin and PCB content in a fish. The link function used is logarithmic and adaptive Gauss-Hermite quadrature is used to approximate the likelihood. Where in a body of water a fish is caught is treated as a random effect. Results indicate that fat content, length, and in which body of water a fish is caught is of importance. Seasonal changes and spatial changes within a water is indicated to be of less importance.

Acknowledgements

I would like to express my gratitude toward Hans Nyquist for his dedication in providing support with every aspect of writing a thesis in statistics; IVL Swedish Environmental Research Institute, especially Magnus Karlsson and Joakim Hållén, for their helpfulness and contribution with knowledge in the field of water and environmental engineering. I would also like to thank Gunnar Andersson from the Swedish Veterinary Institute for his contribution.

Contents

1 Introduction

This thesis constitutes a part of a larger project called ”Dioxins in oily fish- threats and potential for development of small scale coastal and lake fishing” by IVL Swedish Environmental Research Institute, SLU Aqua and The Swedish Veterinary Institute. The project is running up until 2019 and the main purpose is to increase knowledge about how levels of dioxin-like compounds (DLCs) varies over time and space, and thus be able to draw conclusions in order to optimize small scale professional fishing, with respect to levels of DLC in fish going for sale. The bodies of water included in the project are the Swedish coast of the Gulf of Bothnia, Lake Vänern and Lake Vättern. Several species are included in the program, but in this thesis, focus will be on European whitefish (*sik* in Swedish).

1.1 Questions for this project

As the situation is today, prevalence of DLCs in Swedish aquatic ecosystems limits the potential to develop small scale fishing (Karlsson et al. 2018).

Regarding European whitefish, the Swedish National Food Agency has started a program for food control in Lake Vänern and Lake Vättern with respect to dioxin-like substances. This implies that every fisherman has to assure that a catch passes EU-regulations before it can go for sale. Performing laboratory analyses on every batch of fish to investigate DLC content would be very expensive and make sale of fresh fish impossible. Therefore, alternative ways of assuring that DLC content is low are investigated. As a part of investigating this, it is of interest to increase knowledge about DLC content in European whitefish and to use this knowledge when discussing how small scale professional fishing should be regulated. A comment is that species like herring, salmon and trout are not subject to a food control program, only recommendations for yearly consumption.

Here, statistical modelling is applied to data produced by the project in order to investigate if a systematic way of fishing that reduces content of DLCs in European whitefish can be applied. For this thesis, interesting questions are:

- How does morphometric, spatial and temporal changes affect levels of DLCs in European whitefish in Lake Vänern, Lake Vättern and along the Swedish coast of the Gulf of Bothnia.

Morphometric measures are bodily features of a fish, such as length and weight. Spatial variations are variations within a body of water, as well as between different waters. Temporal changes will be seasonal and yearly.

- Are previous believes enhanced?

Ideas about what affects DLC content in European whitefish have emerged during the project ”Dioxins in oily fish- threats and potential for development of small scale coastal and lake fishing”. It is of interest to see if these are justified or if something contradictory is found. The question about previous believes relates to the first question. There exist ideas in what direction, and to what magnitude morphometric, spatial and temporal changes affects DLC content. This will be clarified later on and discussed in the discussion and results section.

In an attempt to answer the research-questions, applying statistical methods and techniques is reasonable for a number of reasons. First of all, statistical procedures are well developed for situations similar to this, in which it is of interest to search for association between levels of a DLC measure (dependent variable) and explanatory variables. Also, an appropriate statistical analysis with formal tests gives a way to quantify when differences of some quantities are large enough to be regarded as statistically significant or small enough to be reasoned to appear by chance. These properties justifies construction of a model consisting of a dependent variable expressed as a linear combination of some predictors. Also, linear models are very flexible and can be adapted to many situations that may appear when wanting to investigate how some factors are associated with a response variable. Also, techniques for model evaluation are well developed.

A comment is that the aim for this thesis is not to extrapolate results to other waters, or other species, and thus, no emphasis will be put on that. The aim is to increase knowledge about what affects DLC levels in European whitefish in Swedish waters, and use it as a contribution when forming/discussing a potential control fishing program.

1.2 Dioxins and PCBs

1.2.1 Chemistry and usage

Dioxins are a group of chemicals in which the basic chemical structure is two benzene-rings to were chloride atoms can attach, in different numbers and structures, forming different *congeners*. The number of chloride atoms and in what way they attach to the benzene-ring is related to specific properties of that substance. Dioxin is sub-categorized in two groups: *polychlorinated dibensofurans* called *furans* and *polychlorinated dibenso-p-dioxins*, most commonly called *dioxin*. In this text, when dioxin is mentioned, furans are included and is sometimes denoted as PCDD/F.

PCBs, or *polychlorated biphenyls* also has two benzene-rings onto which chlorine atoms can bind in different numbers and positions, forming different congeners. Some of these congeners are classified as dioxin-like PCBs because of the planar structure the two benzene-rings have, and these are also the most toxic ones. PCBs that are not dioxin-like have an angle between the two benzene-rings, affecting the toxic properties. Due to the molecular similarity between dioxins and dioxin-like PCBs, they show similar toxic properties. Dioxins, furans and dioxin-like PCBs are, due to their similarities, called *dioxin-like compounds* (DLCs).

These substances has historically been used for a variety of purposes in Swedish industries. PCBs main usages has been as an additive in different types of hydraulic fluids and transformer oils, and in some substances used in construction, such as sealants. Using PCBs is prohibited in Sweden since the 1970s, but due to their persistence they are still found in for example fish and sediment.

Dioxins forms as a residue during combustion, such as burning of waste. These emissions enters ecosystems through precipitation. A part of the historical emissions are due to chemical processes in paper- and pulp industries, such as bleaching of paper with elementary chlorine, which was practiced up until the 1990s. Techniques that do not form dioxin as a rest-product has been used since (Hållén and Karlsson 2018). These historical emissions were local, in the sense that they were confined to a smaller area in connection to the industries, in comparison to atmospheric depositions that affect large areas.

1.2.2 Dangers and health issues

Once in an aquatic ecosystem, DLCs are persistent, meaning that they decompose into other chemicals very slowly. Another property DLCs have is that they are fat soluble, leading to them being stored in fat tissue. A consequence of this is that they accumulate to higher concentrations higher up in the food chain.

Animal testing indicates that dioxins and PCBs affects the immune system, reproductive system, hormonal system, and it might be cancerous. High doses can affect the nervous system, and the development of the brain (Cantillana and Aune 2012).

1.2.3 Regulations and recommendations

The European Union has set a limit, with regards to dioxins and PCBs, to when fish is being allowed for sale on the European market. Levels in European Whitefish caught in Lake Vättern are usually under the limit, while levels in Vänern often are above (Karlsson et al. 2018). Current limits are 3.5 pg/g TEQ wet weight for dioxins and 6.5 pg/g TEQ wet weight for dioxins and dioxin-like PCBs combined (Commission 2011). The unit TEQ (toxic equivalents) is discussed in the next section. Wet weight indicates that the laboratory analysis has been made on raw, not dried, muscle tissue.

The Swedish National Food Agency has given a set of recommendations for the amount of consumption of oily fish. Children under 18 and pregnant women are recommended to eat oily fish from the Baltic sea, Vänern and Vättern at most 2-3 times per year. Others are recommended not to eat such fish more than once a week (Cantillana and Aune 2012).

1.2.4 Toxic Equivalents

There are 210 congeners of PCDD/Fs and 209 of PCBs. In order to quantify the toxicity, the World Health Organisation (WHO) derived a system where the concentration of each congener in a sample is multiplied with a toxic equivalency factor (TEF) and added together to form a measure of toxic equivalents (TEQ). The weights (TEFs) are expressed in relation to the most toxic congene, 2378-TCDD which has $TEF = 1$. For less toxic congeners the value is between zero and one. TEQ allows for obtaining a single measure of how toxic a sample is (Van den Berg et al. 2006).

1.2.5 Previous findings

Some of the previous findings in the project and other reports helps to give a clearer picture of the situation. It is well established that fat content in fish has a positive correlation with levels of DLCs (Karlsson et al. 2018). An idea to use fat content as a proxy for DLC content in European whitefish has been tested, but results when using a hand-held device to measure fat content of fish show much variation. Reliability need to be increased (Karlsson et al. 2018).

The long term trend is that levels of DLCs are reducing in Swedish biotas, although in recent years, the trend for PCDD/F is not as strong as for earlier years (Malmaeus and Karlsson 2014).

Along the Swedish Coast of the Gulf of Bothnia, historical emissions from industries, now stored in sediment, seem to have a local effect on levels in fish, (Malmaeus, Karlsson, and Rahmberg 2012). For some congeners of PCB in perch, concentrations were shown to be higher closer to urban/industrial areas (Nyberg et al. 2014).

Regarding European whitefish, variation in levels of DLCs seem to differ between waters. In Lake Vänern, levels seem to be higher, in comparison to Lake Vättern and the Gulf of Bothnia (Karlsson et al. 2018).

2 Data

Data used for the analysis are supplied by IVL Swedish Environmental Research Institute and are part of the project "Dioxins in oily fish- threats and potential for development of small scale coastal and lake fishing" mentioned earlier. Data are collected between 2013 and 2019. Fish are caught either by staff from IVL or from local fishermen in Lake Vänern, Lake Vättern or the along the Swedish coast of the Gulf of Bothnia (a few observations of European whitefish has been caught in Denmark and the Netherlands, but these are not considered here). Whole, frozen fish has been delivered to IVL:s laboratory in Stockholm where preparation of fish has been made in accordance with EU-regulations and instructions from the Swedish National Food Agency. Information such as date of catch, location of catch etc are supplied by the fishermen. Final laboratory analyses to obtain fat content and concentrations of DLCs have been made by the laboratories ALS (Praha, Czech Republic) and Eurofins (Hamburg, Germany). Species included in the program are European whitefish, herring, trout, salmon, perch and pike, but as mentioned, focus in this study is on European whitefish, of which there are 268 observations.

2.1 Possible influences on conclusions

The procedure of collecting data affect the way the model is constructed. For example, fishermen have been reporting themselves were they caught the fish. They have not followed a set scheme of where to catch the fish, and fish have been caught at many different locations (38 locations to be more precise). Also, a body of water can be seen as an area in which there exist a very large number of possible sites to catch a fish. Since fishermen catch fish where they want, from a choice of many possible catch-sites, this makes, in statistical terms, the catch-sites being seen as randomly chosen from a larger population of catch-sites. To use all the information available, and to avoid the unwanted scenario of using dummy-variable-coding to describe it, the parameter associated to catch-site can be treated as a random variable, in contrast to being treated as fixed. This leads to conclusions being drawn in terms of how much variation the catch-sites contribute with, rather than in levels of DLC for every site. An alternative would have been to set up a fishing-scheme with a smaller number of predetermined locations, in which locations could have been treated as fixed effects, or to condense the full information by re-coding it to have fewer levels, making a dummy-coding more suitable. Similarly, during what time of year fish are caught is not predetermined by a scheme, having lead to no European whitefish being caught during summertime in Vänern. This is clearly not optimal, but it has to be mentioned that it is common practice not to fish European Whitefish in the summertime i Vänern, thus there are no observations for this time of year. Furthermore, conclusions about a season-variable might be difficult to interpret if it turns out to be statistically significant. The main theoretical belief about a season effect is that when the fish spawn, they release the rum, which is high in fat, and since DLC are fat soluble, some amount leave the fish with the rum. The issue is that the population of European whitefish consist of different sub-species that spawn at different seasons, and what sub-species a fish belongs to is not reported. This means that a possible conclusion about a season effect might not be applicable in general, since it can depend on a variable that is not observed.

Also, there are substantial amounts of missing data for some variables that could have been a part of the final model. An example of this is that many fishermen have excluded to report with what method they caught the fish. Fishing method could have been reasonable to include in a model. Missing data reduces the possibility to construct a more extensive model, and thus being able to evaluate the necessity of certain variables. It could have been the case that fishing-method was irrelevant and could be dropped from a model, but having the possibility to include it is preferable.

This relates to construction of linear models in general. If some important variable is excluded, it could result in a mis-specified model that is not able to capture the underlying structures in data, possibly resulting in poor inference, being noticed by structures in residuals.

The structure of collected data might affect conclusions drawn in the sense that it might reduce the possibility to find associations between variables. One thing that supports a case when potential relations that exist can not be found is that data are highly unbalanced. For example, the number of fish caught at a location varies a lot between locations. For some locations, around 20 fish have been caught, and for some, only one. Also, the number of observations per season is unbalanced. From a perspective of statistical power, this is not optimal. Given a certain amount of observations, a balanced design increases the possibility to detect associations that do exist, in contrast to a more unbalanced design.

Another way in which the process of collecting data might influence conclusions is the fact that collective samples have been performed for some of the observations. Collective sample means that muscle tissue from a number of fish caught at the same time and location have been mixed together and then analyzed, resulting in a single measure of fat content and concentrations of DLCs for a number of fish. That measure is therefore an average, and in the corresponding record in the data, measurements such as length and weight are also averages. This procedure has been done to reduce costs since laboratory analyses are expensive. The distribution of collective samples is seen in Table 1.

#individuals	#records
1	204
2	12
3	8
4	7
5	4
6	3

Table 1: Frequency table of collective samples of European whitefish

By taking averages and not obtaining data for all fish, the amount of available information is reduced, and the total variance in data is reduced as well. Having information about all fish, as compared to averages for some, is preferable. In an analysis, an approach to deal with these collective sample must be chosen. One possibility is to include an observation from a collective sample as if it is just one observation, not taking into account that the observation in fact is an average. Another approach is to use weights that correspond to the number of fish in the collective sample. If a collective sample was performed on three fish, that record could be given a weight equal to three. Simulations performed indicates that both these approaches leads to hypothesis tests being more conservative (for code, see Appendix). A third approach is to exclude the observations that are collective samples, but it leads to disregarding a lot of information, and is therefore not considered a viable option. By looking at Table 1 it is clear that most observations are individual samples, and that most collective samples are made on a few number of fish. This leads to a belief that the collective samples is not a major problem when interpreting estimates. The chosen approach to handle these is to use the number of individuals as weights. A collective sample contains more information than for only one fish, and including weights are reasoned to be the ”least bad” way to handle the lack of full information.

2.2 Response variable

When it comes to modelling, a measure of DLC content will be the response variable.

For every analyzed individual, data include concentrations for 17 congeners of PCDD/F, 12 congeners of dioxin-like PCB and 6 PCB congeners that are not dioxin-like. In this thesis, analysis will be made using the one measure considered to contain the most information about DLCs in European whitefish. This is the sum of all 29 (12+17) dioxin like-compounds weighted with their corresponding TEF-value, resulting in a single TEQ-value of PCDD/F and PCB content. Formally, this is denoted $\sum PCDD/F + dl - PCB - TEQ$ *pg/g ww*, but is for efficiency here referred to as *dioPCB*. The reader is reminded that in general, DLC content is measured as concentrations, and *dioPCB* is a measure of toxicity, related to the composition of congeners in a sample.

Other potential dependent variables, such as $\sum PCDD/F - TEQ$ *pg/g ww*, $\sum dl - PCB - TEQ$ *pg/g ww* or the concentration of the PCBs that are not dioxin-like is also of interest for the project ”Dioxins in oily fish- threats and potential for development of small scale coastal and lake fishing”, but is not considered here.

The distribution of *dioPCB* is seen in Figure 1.

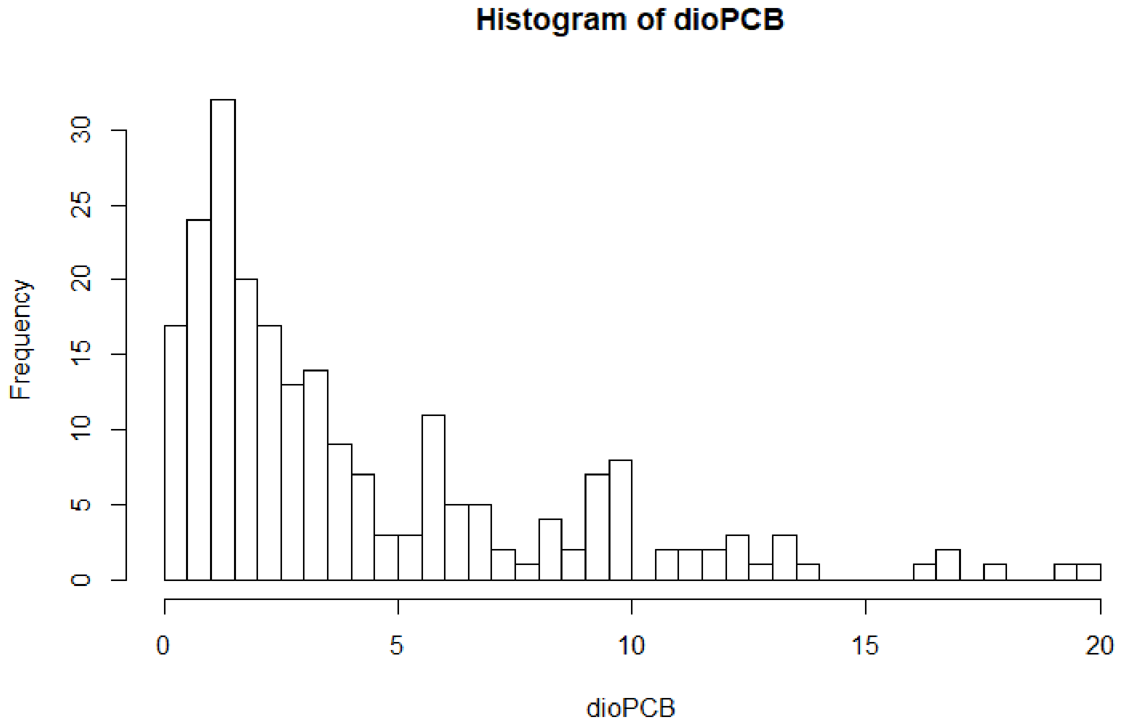


Figure 1: Dioxins and dioxin-like PCBs in analyzed samples of European whitefish.

2.3 Explanatory variables

Available explanatory variables that are candidates to be included in a model are seen in Table 2. The variables ”Year,” ”Season” and ”Binary.Length” are not included in original data provided by the project but are created from existing variables. ”Season” and ”Year” was created using information from ”Date of catch” and ”Binary.Length” is a dummy-coding of the original ”Length of the fish” variable, taking value zero if smaller than 38 and one if larger. The reason is that a potential part of a control-program for sales of European whitefish is to only allow sales of fish shorter than some specified length. A proposal has been set to 43 cm, but data do not support testing this limit since it contains very few fish longer than 43 cm, so for a more balanced binary variable, the limit here is set to 38 cm. All tests for European whitefish was performed by ALS (Praha,Czech Republic), so there is no need to include *Laboratory* as a blocking variable. ”Year” is treated as a continuous variable.

European whitefish has been caught at a total of 38 different locations (catch-sites) in Lake Vänern, Lake Vättern and along the Swedish coast of the Gulf of Bothnia. As mentioned previously, the number of observations per location is unbalanced (Table 3). To get a perspective of how these are distributed, they are shown for Lake Vänern and Vättern (Figure 2). Catch-sites for the Gulf of Bothnia are not included in the figure but they are along the Swedish coast and ranging from Tierp in Uppland to Torne in the most northern part.

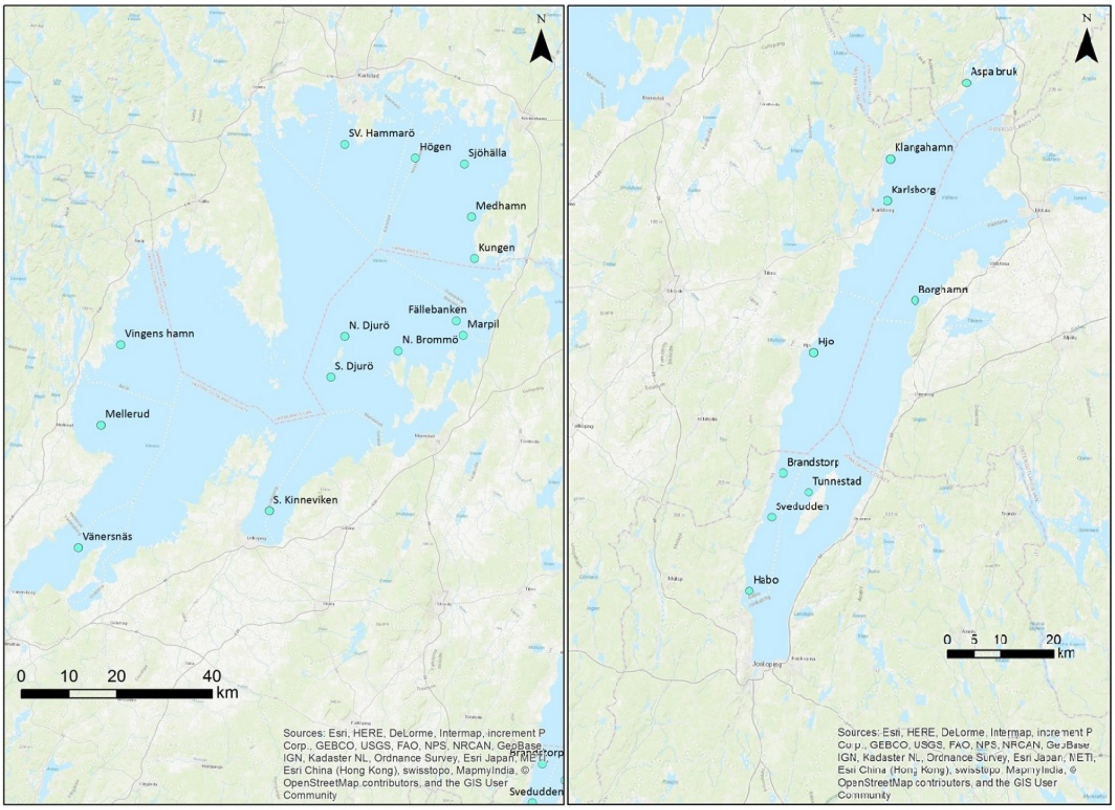

N
SV. Hammarö
Högen
Sjöhälla
Medhamn
Kungen
Fällebanken
N. Djurö
Marpil
N. Brommö
Vingens hamn
S. Djurö
Mellerud
S. Kinneviken
Vänersnäs
Brandstorp
Svedudden
0 10 20 40 km
Sources: Esri, HERE, DeLorme, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), swisstopo, MapmyIndia, © OpenStreetMap contributors, and the GIS User Community
N
Aspa bruk
Klangahamn
Karlsborg
Borghamn
Hjo
Brandstorp
Tunnestad
Svedudden
Habo
0 5 10 20 km
Sources: Esri, HERE, DeLorme, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), swisstopo, MapmyIndia, © OpenStreetMap contributors, and the GIS User Community

Figure 2: Catch-sites of European whitefish in Lake Vänern and Lake Vättern. Gulf of Bothnia not included but catch-sites there are along the Swedish coast and ranging from Gräsö located outside Tierp to Torne in the north.

Variable name	NAs %	categorical/continuous	#levels
Laboratory	0	Cat	1
Body of water	0	Cat	3
Site of catch	0	Cat	38
Date of catch	4.2	-	-
Year	4.1	Cont	-
Season	4.2	Cat	4
Date of preparation*	10.9	-	-
Coordinates of catch	21.8	Cont	-
Number of individuals*	0	Cat	6
Sex	12.3	Cat	2
Binary.Length	1.26	Cat	2
Length	1.26	Cont	-
Weight	10.9	Cont	-
Somatic weight*	0.42	Cont	-
Liver weight	19.3	Cont	-
Gonad*	24.8	Cont	-
CF (whole)*	10.9	Cont	-
CF (Somatic)	1.68	Cont	-
LSI*	8.82	Cont	-
GSI*	14.3	Cont	-
Fat percentage	0.84	Cont	-
Fishermen	37.4	Cat	14
Fishing method	52.6	Cat	4
Catch depth	56.7	Cont	-
Age(otolith)*	70.2	Cont	-
Age(scale)*	77.7	Cont	-

Table 2: Possible explanatory variables to use in a model.*Date of preparation=when muscle tissue was prepared for analysis. Number of individuals = some catches of fish was analyzed as collective samples, meaning that one record represents averages from a number of fish. Somatic weight is the weight of gutted fish. Gonad is the weight of the reproductive organs in the fish. Condition factor describes the body shape of the fish, calculated using length and weight, whole or somatic. LSI= Liver Somatic Index, is the ratio of the liver weight to the somatic weight of the fish. GSI = Gonad Somatic Index, is the ratio of gonad weight to the somatic weight of the fish. Otholith is a calcium carbonate structure in the inner ear. Its growth rings allows for determining age. Age is also determined by examining fish scales.

#individuals	#Locations
1	12
2	5
3	3
4	4
5	2
6	2
7	2
9	1
10	1
11	1
14	1
17	1
18	1
24	1
38	1

Table 3: Frequency table of number of fish per catch-site.

3 Method

3.1 Motivation of using a generalized linear model

First of all, it can be of interest to motivate why a usual linear regression model without any variable transformations is not a good choice for the intended analysis, and then to motivate a choice among the remaining alternatives.

The most important features of the dependent variable here is that it is continuous, values can not be negative and most observations have a small value.

A lower bound at zero is not in itself an issue in linear regression but it can be. Let's imagine a continuous variable with a lower bound at zero, but with a large mean value, low standard deviation, and a fairly even distribution of observations around the mean. Linear regression will most likely work well for such a dependent variable and given that the linear predictor is specified well, inference will be credible. For *dioPCB*, many observations are small and positive, leading to a skewed distribution. Performing linear regression will now violate the assumption of normally distributed error terms, which is bad for inference.

Also, regarding the assumption of homoskedasticity, observations closer to the boundary are likely to show less variability, which could result in a clear violation of the assumption. To summarize, a linear regression without any transformation is not preferable since it leads to violations of the assumptions of homoskedasticity and normality of error terms.

To resolve these issues there are two main strategies. One is to apply a transformation to the dependent variable, and the other is to use a generalized linear model (GLM).

Applying a transformation to the dependent variable, often a logarithmic or square root, could result in less violation of the assumption of normally distributed residuals. If homoskedasticity still is violated and the structure of how variances increases is known, it can be accounted for using weights. An advantage with applying a transformation is the simplicity, since the same procedures as for linear regression can be used. A downside can be difficult interpretation. Interpretation is easily made in terms of the transformed variable, but can be more demanding in the original scale of the variable.

Regarding the option of using a generalized linear model, a distribution of the dependent variable has to be chosen. As mentioned, *dioPCB* is continuous and bounded below by zero, so a distribution with these characteristics is suitable. Although, when relating this to the discussion about increasing variance with increasing mean, the best choice is to assume a gamma-distribution. The reason is that the gamma distribution has a constant coefficent of variation (CV) that takes the increasing variability into account (Faraway 2016). In section 3.11 the calculation of the CV is shown. By using a gamma-GLM, both the problems of homoskedasticity and normality of error terms are taken into account. It is shown in section 4.3 that the variance does in fact increase with an increasing mean. Interpretation of a generalized model can be more complicated, or it can be similar to a linear regression, depending on the choice of what is called the *link-funtion*. This is explained more in depth in section 4.4. For now it is concluded that multiple choices that might be more or less suitable exist. A downside with using a gamma-GLM instead of applying a transformation and assuming a variance structure is that most examples regarding model evaluation for GLMs are based on other distributions, often binomial or Poisson. Since model evaluation differs for different distributions, it is easier to find information about model fitting and model evaluation when using a linear regression model.

With the discussion about assumptions and being able to choose between different link functions in mind, a generalized linear model assuming a gamma distributed dependent variable is regarded as the best choice. To be more specific, a generalized linear mixed model is used. It is explained later on why it is reasonable to use a model where some model parameters are treated as fixed, and other as random. If a transformation would have been the choice, and a random parameter added, the model would have been called a linear mixed model. The choice between generalization or transformation is based on the same argumentation if random effects are included in the model or not.

3.2 The concept of generalized linear mixed models

A generalized linear mixed model is in essence an extension of a mixed effects model into the framework of generalized linear models. For clarity, the theory section in this thesis is therefore comprised of several parts. They introduce and describe the fundamentals of mixed effects models, generalized linear models, merges the concepts into generalized linear mixed models and describes difficulties with estimating these types of models. Described procedures are put in relation to a gamma distributed dependent variable.

3.3 Fixed effects

A mixed effects model can be seen as a combination of two commonly used techniques; the fixed effects model and the random effects model.

The fixed effects model only includes explanatory variables that are regarded as non-stochastic. The error term is the only random component of the model, and the explanatory variables are either categorical or numerical. In matrix notation, a fixed effects model can be expressed as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \boldsymbol{\varepsilon} \sim N(0, \mathbf{R}), \tag{1}$$

where $\boldsymbol{X}$ is the design matrix, $\boldsymbol{\beta}$ is the parameter vector and $\mathbf{R}$ is the covariance matrix of the error terms. If the assumption of independence holds, all off-diagonal elements of $\mathbf{R}$ are equal to zero. In practice that is never the case, and residual analysis is used as a tool for indicating if the deviation from the assumption is large, in which case inference is poor.

3.4 Random effects

In an study, if factors levels for which measurements are made are chosen at random from a larger population of levels, possibly infinitely large, then it can be suitable to regard this effect as a random effect. This is often the case in biology and ecology. In some cases, as for repeated measures designs, subjects on which repetitions are made can be seen as a grouping variable that is random. It is usually assumed that the random effect is normally distributed with mean zero and some variance. By treating a factor as random, estimation is not made for specific factor levels, but rather the amount of variability the factor contributes with. Another reason for treating a factor as random is that it might have many levels, and treating it as fixed leads to estimation of many parameters, possibly resulting in an overfitted model.

Random effect models are specified in the same manner as the fixed effects, the difference being that the components of the parameter vector are treated as normally distributed random variables.

3.5 Mixed Effects Model

If a model includes both fixed and random effects, it is called a mixed effects model. Below, the structure of a mixed model is specified, and the assumptions are stated:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon} \quad \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim N\left(0, \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix}\right), \tag{2}$$

where $\boldsymbol{X}$ is the design matrix for the fixed effects, $\boldsymbol{\beta}$ is the parameter vector for the fixed effects, $\boldsymbol{Z}$ is the design matrix for the random effects and $\boldsymbol{u}$ is the parameter vector for the random effects.

It is worth mentioning that assuming multivariate normality of the random components is not necessary, but it is often done in order to make inferences based on normality. Zero mean and no correlation between $\mathbf{u}$ and $\boldsymbol{\varepsilon}$ on the other hand is a stricter assumption. Estimated model parameters are maximum likelihood estimates. Important is that maximum likelihood estimates are only asymptotically unbiased, so in small sample situations the restricted maximum likelihood (REML) is an alternative (Lindstrom and Bates 1990). Both maximum likelihood and restricted maximum likelihood estimates of parameters can be solved for by a closed form expression, so an iterative procedure is not needed. Models with more than one random effect have some structure of the random effects, nested or crossed, but since it will not appear in this thesis it is left unexplained. More information can be found in for example (Montgomery 2017, Ch.13). Here, one random effect is included in the model, and it is said to add a *random intercept.* For every observation, a random component associated with the corresponding level of the grouping variable is added to the usual intercept. The grouping variable here is the location within a body of water where a fish is caught (the catch-site).

3.6 Generalized linear models

Generalized linear models (GLM) is a family of models used when we want to fit a linear model, but the response variable is not normally distributed. It was proposed by Nelder and Weddeburn in (Nelder and Wedderburn 1972) and is used extensively. Common usages is when response distributions are binomial, Poisson and of gamma-type . A GLM has three main components, stated below:

- A response variable Y which follows a distribution that is an exponential family.
- A linear predictor $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$ where $\mathbf{x}_i$ is one row of the design matrix and $\boldsymbol{\beta}$ is the parameter vector.
- A link function g such that $g(E(Y_i)) = g(\mu_i) = \eta_i$ (Dobson and Barnett 2018).

Expressed in words it can be said that the GLM-procedure model expected values of the response, subject to a link function, as a linear combination of explanatory variables. A computational difference between linear regression models and generalized linear models is that maximum likelihood estimates of the parameters has to be solved iteratively for GLMs. Another difference is that for the GLM, expected values are modelled, so no error term is included in the linear predictor, whereas in linear regression models the observed values are modelled, leading to the error term being included in the linear predictor. By giving a short example it is clear that the usual linear regression model is a special case of a GLM: if the link function g is the identity-link, meaning that $g(\mu_i) = \mu_i$, and Y_i is assumed to follow a normal distribution, the model is a linear regression model.

3.7 Gamma distribution and Exponential family

A probability distribution is an exponential family if it can be written on the form:

$$f(y;\theta) = e^{a(y)c(\theta)+b(\theta)+d(y)} \tag{3}$$

The functions $a(.)$, $b(.)$, $c(.)$ and $d(.)$ form the basis for constructing the equations for parameter estimation in a GLM. Details of that process will not be shown here, but is clearly stated in (Dobson and Barnett 2018). If the function $a(y) = y$ the distribution is said to be in canonical form, and if in addition, $c(\theta) = \theta$, θ is said to be the *canonical parameter.* If also, a dispersion, ϕ, is included in the definition (otherwise, the dispersion is included in other functions), the definition in (McCullagh and Nelder 1989) is obtained:

$$f(y|;\theta,\phi) = exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + d(y,\phi)\right) \tag{4}$$

If a distribution has more than one parameter, all but one has to be treated as fixed. That is the case in this thesis where responses are assumed to follow a gamma distribution, which is an exponential family, and has the following density function:

$$f(y;\alpha,\gamma) = \frac{\gamma^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\gamma y} \quad 0 < y < \infty \quad \alpha, \gamma > 0,$$

with

$$E(Y) = \frac{\alpha}{\gamma},$$

and

$$Var(Y) = \frac{\alpha}{\gamma^2}$$

In order use to this distribution in a GLM, α is treated as a constant, so it is a function of the dispersion parameter and γ is related to θ.

3.8 Generalized linear mixed models

In this section, the GLM-methodology is applied to the linear mixed model. These type of models arise due to two characteristics: Assuming a non-normal response and the inclusion of one or more random effects. The inclusion of random effects makes data being viewed as grouped or clustered. In, for example repeated measures designs, the same subject is believed to have correlated outcomes for the different repetitions, and in this thesis, when catch-site is included as a random effect, fish caught at the same location leads to a belief of DLC levels in these observations being correlated instead of independent. Observations within the same group are not seen as independent, but independent when conditioned on the grouping variable (the random effect). In matrix notation when the number of random effects is not restricted, applying the link function to the conditioned Y:s yields:

$$g(E(\boldsymbol{Y}|\mathbf{u})) = g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Zu} \tag{5}$$

When only one grouping variable is included, it introduces an indexation where ij denotes the i:th observation within the j:th level of the grouping variable. Each level of j has n_j observations and there are N levels. For a generalized linear model with one random effect, the linear predictor is:

$$\eta_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + u_j, \tag{6}$$

where $\boldsymbol{x}_{ij}^T$ is the row in the fixed effects design matrix corresponding to observation ij, $\boldsymbol{\beta}$ is the fixed effects parameter vector and u_j is the random component associated to the j:th level of the grouping variable.

The main issue with parameter estimation in generalized linear mixed models is that the usual maximum likelihood or restricted maximum likelihood approach does not work. Due to the inclusion of random effects in the linear predictor, the likelihood consists of an integral that can not be evaluated analytically unless a normal distribution is assumed and the identity link is used (i.e the model is a linear mixed model). A numerical approximation of this integral can be calculated in order to obtain parameter estimates. Another alternative is to use a Bayesian approach.

Now the likelihood is specified. Since the observations in $\mathbf{y}$ are assumed independent only when conditioned on the random effects $\mathbf{u}$, and the likelihood of interest is the marginal likelihood of $\mathbf{y}$, the joint density has to be integrated over the random effects. The full likelihood is given by:

$$\begin{aligned} L(\boldsymbol{\beta}, \phi, \mathbf{G}; \mathbf{y}) = \int f(\mathbf{y}, \mathbf{u} | \boldsymbol{\beta}, \phi, \mathbf{G}) d\mathbf{u} = \\ \int f(\mathbf{y}|\mathbf{u}; \boldsymbol{\beta}, \boldsymbol{\phi}) f(\mathbf{u}|\mathbf{G}) d\mathbf{u} = \\ \prod_{j=1}^{N} \int \prod_{i=1}^{n_j} f(y_{ij}|\mathbf{u}_j; \boldsymbol{\beta}, \boldsymbol{\phi}) f(\mathbf{u}_j|\mathbf{G}) d\mathbf{u}_j \end{aligned}$$

In relation to data used here, a random intercept model is used, so $\mathbf{u}$ is not a vector and the variance-covariance structure of $\mathbf{G}$ is a scalar, denoted τ^2. The likelihood then becomes:

$$L(\boldsymbol{\beta}, \phi, \tau^2; \mathbf{y}) = \prod_{j=1}^{N} \int \prod_{i=1}^{n_j} f(y_{ij}|u_j; \boldsymbol{\beta}, \boldsymbol{\phi}) f(u_j|\tau^2) du_j \tag{7}$$

3.9 Method of approximating the likelihood function

As mentioned, the likelihood function in a generalized linear mixed model can not be evaluated analytically, resulting in an approximate method having to be chosen. A variety of approaches have been suggested and implemented in software for this purpose. These include the penalized and marginal quasi-likelihood (Breslow and Clayton 1993), maximum likelihood approaches such as the Monte Carlo EM-method and Monte Carlo Newton-Raphson (McCulloch 1997) and Bayesian approaches such as the Markov-Chain Monte

Carlo approach (Hadfield 2010). Maximum likelihood-based methods relies on numerical approximations of the likelihood, often by using Gauss-Hermite quadrature, adaptive Gauss-Hermite quadrature or Laplace integration.

However, studies of comparisons between these approaches are limited, and is potentially an area for further research, especially for the case when the dependent variable is assumed to be gamma-distributed. Most comparisons of methods are based on data sets in which the dependent variable is binary, or a count. No papers of comparisons between methods when performing a generalized linear mixed gamma model have been found. This results in a choice being made based on the more general findings that has been found. For example, in (Rabe-Hesketh, Skrondal, and Pickles 2002) it is argued that advantages of marginal quasi-likelihood and penalized quasi-likelihood is computational efficiency, the downside is that they perform poor for binary data with small cluster sizes. MCMC-methods seem to be a reasonable alternative, but downsides being that they are computationally intensive and that it can be difficult to evaluate when the chain has converged to a stationary distribution. Gauss-Hermite quadrature appear to be a reasonable alternative. By adjusting the number of quadrature points, exactness of approximations increases but the process becomes more computationally intensive. Gauss-Hermite quadrature also gives values for likelihoods which can be a basis for likelihood-ratio tests (Rabe-Hesketh, Skrondal, and Pickles 2002). Downsides appear to be possible poor performance in some situations involving binary or Poisson data, when intraclass correlation is high. Another downside is that inclusion of many random effects in combination with many quadrature points makes Gauss-Hermite quadrature computationally intensive. Adaptive Gauss-Hermite quadrature is a modification of the ordinary Gauss-Hermite quadrature that yields a better approximation but increases computational intensity even more. In terms of computational intensity of standard and adaptive quadrature, the main increase is due to inclusion of many random effects in the model.

With this discussion as a basis, the choice is to use adaptive Gauss- Hermite quadrature. In models in this thesis, only one random effect is included, making the procedure not so computationally intensive, and the ability to change the amount of quadrature points in order to achieve a better approximation is regarded as an advantage. Also, likelihood-ratio tests can be of interest when evaluating models. The downsides that have been discovered do not seem directly applicable to data used in this thesis. Gauss-Hermite is implemented as a standard procedure in both R and SAS, which could serve as an informal indication of strengths of the method. The *R*-function *glmer* is constructed based on adaptive Gauss-Hermite quadrature and will be used for estimation of models. This program uses a maximum likelihood approach in combination with approximating the intractable integral using adaptive quadrature. As seen in (McCulloch 1997), several ways to implement a maximum likelihood approach exists, and descriptions of the *lme4*-package of which *glmer* is a part of, do not specify exactly which maximum likelihood approach that is implemented.

3.10 Gauss-Hermite quadrature

The basic idea of Gauss-Hermite quadrature is to approximate an integral by a summation consisting of function values and weights. In its most general form, Gauss-Hermite quadrature can be written as:

$$\int_{-\infty}^{\infty} f(x)e^{-x^2}dx \approx \sum_{j=1}^{N} w_j f(x_j) \tag{8}$$

Where w_j is a weight and N is the number of points x_j (quadrature points) to evaluate the function $f(x)$ at.

A re-expression of the integral in (8) that is more relatable to the likelihood in (7) allows for an arbitrary Gaussian density instead of just e^{-x^2} and is expressed as:

$$\int_{-\infty}^{\infty} f(x)\phi(x;\mu,\sigma^2)dx \approx \sum_{j=1}^{N} w_j^* f(x_j^*)$$

Where $\phi(x;\mu,\sigma^2)$ is the Gaussian density. The weights w_j^* and nodes x_j^* are transformations of the weights and nodes corresponding to (8).

In terms of a generalized linear mixed model, $f(x)$ corresponds to the conditional density $f(y|u)$ and the Gaussian density corresponds to $f(u)$ since $f(u)$ is assumed to follow a normal distribution.

The quadrature points x_j and weights w_j are derived from the Hermite polynomials, which is a sequence of polynomials of increasing degree, in which any two polynomials are orthogonal. If N quadrature points are chosen, the $x_j : s$ are the N roots of the $N : th$ degree Hermite polynomial, and the weights are a function of the $x_j : s$. In the case of Adaptive Gauss-Hermite quadrature, the quadrature points and weights are modified in order to better approximate the peak of the function to be integrated. It does so by shifting and scaling the quadrature points to locate them under the peak of the function. If the number of quadrature points equals one, the Adaptive Gauss-Hermite quadrature is equivalent to a Laplace approximation, which is a simpler but common technique for numerical integration. A more thorough explanation can be found in (Rabe-Hesketh and Skrondal 2004).

3.11 Basics of Generalized Linear Gamma Models

Using the function in (4), in exponential family theory, the mean μ expressed in terms of the canonical parameter is simply: $b'(\theta) = \mu$ where $b(\theta)$ is called the *cumulant function.* For the gamma distribution, $b'(\theta) = \mu = \frac{d}{d\theta}(-ln(-\theta)) = -\frac{1}{\theta}$. Since gamma distributions has two parameters, as mentioned previously, α is treated as constant and can be shown to equal $\phi = \frac{1}{\alpha}$, called the *dispersion parameter.*

Note that the gamma-GLM can handle dispersion in a good way. When treating α as a constant, the coefficient of variation (CV) is constant:

$$CV = \frac{\sqrt{Var(Y_i)}}{E(Y_i)} = \frac{\sqrt{\alpha \gamma_i^{-2}}}{\alpha \gamma_i^{-1}} = \frac{1}{\sqrt{\alpha}}$$

After fitting the model, the maximum likelihood estimate of the dispersion parameter can be obtained using:

$$\hat{\phi}_{ML} = \frac{1}{\hat{\alpha}} = \frac{\chi^2}{n-p}$$

where χ^2 is the Pearson χ^2-statistic, $\chi^2 = \sum \frac{(y_i - \hat{\mu_i})^2}{V(\hat{\mu_i})}$, n is the number of observations and p is the number of parameters in the model. The estimate of the dispersion parameter is not always of interest when evaluating models.

3.11.1 Deviance residuals

For generalized linear models, checking normality for the residuals $y_i - \hat{\mu_i}$, where y_i is an observation of the dependent variable and $\hat{\mu_i}$ is the estimated mean, is not as meaningful as for linear regression. Instead, *deviance residuals* are a more suitable measure. They are the contribution from each data point to the deviance measure specified as:

$$D = 2(l(\hat{\boldsymbol{\beta}}_{max}|\mathbf{y}) - l(\hat{\boldsymbol{\beta}}|\mathbf{y})),$$

where $l(\hat{\boldsymbol{\beta}}_{max}|\mathbf{y})$ is the likelihood for a model with the maximum number of parameters that can be estimated, a *saturated model.*
For a generalized linear gamma model, the deviance is:

$$D = -2\sum_{i=1}^{n}\left(log\left(\frac{y_i}{\hat{\mu_i}}\right) - \frac{y_i - \hat{\mu_i}}{\hat{\mu_i}}\right),$$

obtaining the deviance residuals d_i as:

$$d_i = sign(y_i - \hat{\mu_i})\sqrt{-2\left(log\left(\frac{y_i}{\hat{\mu_i}}\right) - \frac{y_i - \hat{\mu_i}}{\hat{\mu_i}}\right)},$$

resulting in the deviance being equal to the sum of squared deviance residuals

$$D = \sum_{i=1}^{n} d_i^2$$

3.11.2 The canonical link function

Often i GLMs, the *canonical link* is used as the link function. It is the function $h(.)$ that satisfies:

$$h(\mu) = \theta,$$

and because

$$\mu = -\frac{1}{\theta},$$

the canonical link is the negative inverse, since

$$-(\mu)^{-1} = -(-\frac{1}{\theta})^{-1} = \theta$$

Usually, the minus sign is disregarded, and the canonical link is specified as just the inverse. As mentioned, this is often used as the link function, but other functions can be chosen. For proofs and a more thorough explanation, see (McCullagh and Nelder 1989). Here it is specified explicitly because it gives a simpler expression for the conditional density than another link function would.

When the canonical link is used and the model is extended to a generalized linear mixed model rather than a generalized linear model, y is conditioned on the random component u, and θ is replaced with η, resulting in the following expression for the conditional density:

$$f(y|u;\boldsymbol{\beta},\phi) = exp\left(\frac{y\eta - b(\eta)}{a(\phi)} + d(y,\phi)\right) \tag{9}$$

It is now clear how the linear predictor is a part of the conditional density that is a part of the intractable integral that has to be approximated in order to obtain the likelihood of $\mathbf{y}$.

4 Model Specification and choice of link function

The model specified in equation (6) is a general way to specify a generalized linear mixed model with one random effect. A more specific model is needed for data used in this thesis. Construction of a first model is based on which explanatory variables that are theoretically interesting and reasonable to include (due to missing values) and also on which interactions that are interesting. The view on interaction effects in this thesis is that they should be included based on a theoretical belief that they might be present. First, the model is stated, and then, the reasoning behind it is explained.

The linear predictor is:

$$\begin{gathered}\eta = \beta_0 + u_j + \beta_{1l}Lake + \beta_2 Fat + \beta_3 bin.Length + \beta_4 Year + \beta_5 CFS + \\ \beta_{6k}Season + \beta_{7l}Fat * Lake + \beta_8 Fat * bin.Length + \beta_{9k}Fat * Season\end{gathered} \tag{10}$$

$$l = 1,2 \quad k = 1,2,3 \quad and \quad U_j \sim N(0,\tau^2)$$

where u_j is the random component associated with catch site j.

The coding of fixed effects included in the model is:

- *Lake* corresponds "Body of Water" (although the Baltic Sea is not a lake). Levels are Lake Vänern, Lake Vättern and the Baltic Sea. All areas within the Baltic Sea has been combined to this name, because of few observations in sub-areas.
- *Fat* is fat content. Original scale is percentage of fat in live weight muscle tissue. Scaled to mean zero and unity standard deviation in estimation procedure.
- *bin.Lenght* (binary length) takes the value 0 if a fish is shorter than 38 cm, and 1 otherwise.
- *Year* is treated as continous, with 2015=0, 2016=1 etc. Years includes 2015, 2016, 2017 and 2018. Scaled to mean zero and unity standard deviation in estimation procedure.
- *CFS* is the somatic condition factor, calulated as $\frac{somaticweight}{length^3}$. Scaled to mean zero and unity standard deviation in estimation procedure.
- *Season* is summer, fall, winter or spring.

4.1 Main effects

From variables in Table 2, a linear predictor has been constructed. The aim is to include variables that do not have too many missing values, in order to have as many observations as possible in the final model, and to have both morphometric, spatial and temporal variables. The temporal variables of most relevance are *Season* and *Year.* Some alternatives for morphometric variables are functions of each other, for example *Condition Factor (Whole or Somatic)* is a function of *Length* and *Weight (Whole or Somatic).* The correlation between Length and Somatic Weight is 0.8. The choice is to only include *Length.* The correlation between *Length* and CFS is 0.1 so both can be included without having too high correlation between explanatory variables. The only spatial information not included in the model is the coordinates of where a fish was caught. The name of the location is used instead.

4.2 Interaction effects

The interactions included in (10) are explained and motivated below.

4.2.1 Interaction between length of the fish and fat content

Length serves as a proxy for age here (younger fish are generally shorter). Smaller fish are not top predators, and to a larger extent eat plankton that have lower levels of DLCs as compared to larger fish that have another diet. Because of this, smaller fish can have a lower increase in DLC levels per every unit increase of fat content, as compared to longer fish, that most likely are older. "Age" has too many missing values to be included in the model. A similar argument about DLC levels and age is a basis for the decision by the Swedish government to only allow sales of herring that are smaller than 17 cm in length.

4.2.2 Interaction between "Lake" and fat content

It is believed that the relation between fat content and levels of DLCs might vary between Lake Vänern, Lake Vättern and The Gulf of Bothnia. It is of interest to know if some bodies of water show different degree of association between fat content and levels of DLCs. In the perspective of development of small scale fishing of European Whitefish this interaction might be of small practical use since fat content is difficult to measure, and the proposed control program is only directed towards Lake Vänern.

4.2.3 Interaction between fat content and season

Fat content has previously been seen as a proxy for DLC content, and if season and fat content only are included as main effects, finding a season effect might be difficult since fat content is believed to vary by season and thus leading to change in DLC content. In a regression-type interpretation, a significant season effect is interpreted as "given the fat content (and other explanatory variables), the estimated difference in $dioPCB$ is $\hat{\beta}_x$ between $season = reference$ and $season_k$", but since fat content might change with seasonal changes, an interaction is added.

4.3 Link function

Possible link functions are discussed based on two perspectives. One is how the link function is suitable for data, in terms of model fit and meeting assumptions of the model. The other is how the choice of link function affects interpretability of the model. Best case scenario is that a link function gives a good fit, and parameter estimates that have an intuitive interpretation that is in line with theory about how DLC content in fish is related to different variables.

In general, the criterion for a link function is that it is a monotone and differentiable function. Often, the canonical link is chosen since it has tractable statistical properties, but it is not a necessary choice. For gamma distributions, common link functions are the inverse link $\frac{1}{\mu} = \eta$ (canonical), log link $log(\mu) = \eta$ or identity link $\mu = \eta$. The choice will be one of these. An idea would be to compare the model fit (using some deviance measure) for the same linear predictor and response variable, but with different links. This idea however is not suitable when comparing link functions since models are not nested. A log link for example gives a multiplicative model if exponentiated, while an identity link gives an additive, and deviance measures are not suitable for comparison between such models. Instead, a choice will be based on residual plots and a discussion about which link function that gives a not too complicated interpretation of estimates. First, residual plots are shown and discussed. The model in (10) is fitted with the potential link functions and residual diagnostics are evaluated. Two plots are shown, one with deviance residuals vs predicted values and one with response residuals $(y - \hat{y})$ vs predicted values. In a plot of deviance residuals vs predicted values, two things are of interest: Does the variance change with changing magnitude of the predicted values? Large changes are unwanted since it results in poor inference.

Is there any non-linear pattern in the plot? Any clear pattern in the plot, for example a quadratic, could be an indication of some structure in data that has not been captured by the model. The philosophy adopted when looking at residual plots is that assumptions are always violated, what differs is how severe the violation is.

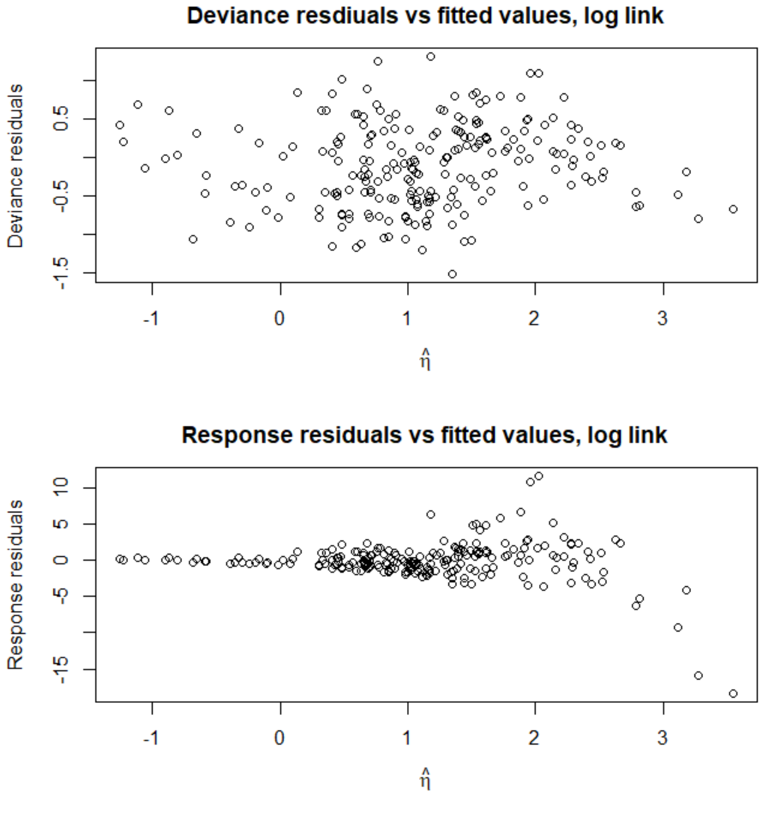


Figure 3: Residuals vs fitted values from estimation of model (10) using log link. Upper plot show deviance residuals vs predicted values. Lower plot show response residuals vs fitted values. Horizontal axis is fitted values in link scale.

From Figure 3, which show plots of residuals vs fitted values for model (10) with log-link, no clear indication of unequal variance is seen in the upper plot. Possibly, variances are slightly lower for high and low predicted values, but there are fewer points here, so it is difficult to see. A pattern indicating non-linearity might be seen. The left and right tail appears to have a slight downward slope. More data could have enhanced or rejected this belief. In general, the plot looks good. No clear violation of assumptions or inability to capture structures in data is seen. The lower plot shows the response residuals vs fitted values. Response residuals have not taken dispersion into account and here it is seen that variances of response residuals increases with increasing fitted values. In fact, this plot acts as a justification of the belief stated earlier that variance increases with increasing mean values. DLC levels was believed to show higher variances for higher levels. This was one of the arguments for choosing a gamma distribution.

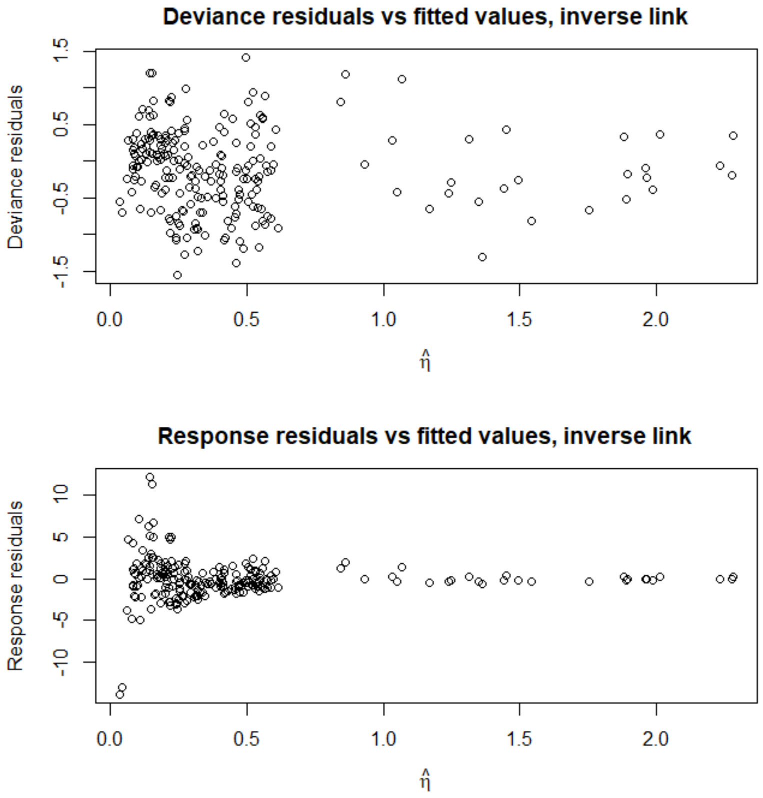


Figure 4: Residuals vs fitted values from estimation of model (10) using inverse link. Upper plot show deviance residuals vs predicted values. Lower plot show response residuals vs fitted values. Horizontal axis is fitted values in link scale.

From Figure 4, which show plots of residuals vs fitted values for model (10) with inverse-link, possibly, lower variance in the right corner is seen, but lower amount of points here makes it difficult to know for certain. No patterns indicating a poorly specified model is visible. In the lower plot, variances are declining for larger fitted values, but since it is in the link scale and the link is the inverse function, variances are actually increasing with increasing mean values, as was stated in the discussion about Figure 3.

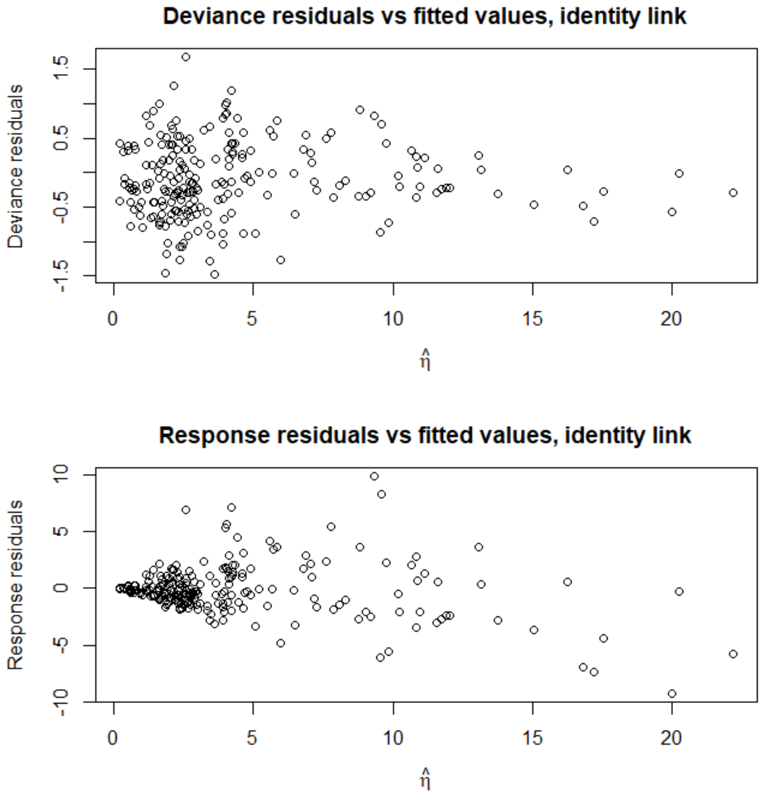


Figure 5: Residuals vs fitted values from estimation of model (10) using identity link. Upper plot show deviance residuals vs predicted values. Lower plot show response residuals vs fitted values. Horizontal axis is fitted values in link scale.

In Figure 5 , which show plots of residuals vs fitted values for model (10) with identity-link, variances seem to be lower for larger deviance residuals, which is not good in terms of meeting assumptions of the model. In terms of patterns indicating non-linearity, none is seen. Regarding the lower plot, variances increase up to a certain point an then seem to even out, which is not in accordance with gamma-theory in which variances always should increase with increasing mean. A general comment for all residual plots discussed is that when it comes to suspicion about patterns and variances, it is the same group of observations that lead to these suspicion. The group is the observations seen in the right hand side of the upper plot in Figure 3, 4 and 5.

4.4 Link function and interpretability

Ideally, a link function gives an interpretation of parameter estimates that agrees with theory on how morphometric, spatial and temporal changes affect DLC content in fish. This model, whose main effects and interactions have been discussed with parts involved in the project, comes from theory similar to a usual regression model, meaning that effects might be believed to be additive, and possible multiplicative associations are dealt with by including interaction terms in the linear predictor. A linear predictor for a gamma model that is in most correspondence with this is the identity link. Using this link gives an interpretation that is the same as for a linear regression. On the other hand, while discussing a possible model with the Swedish Veterinary Institute, a log transformation was proposed to account for skewness of data, so a multiplicative model (in the original scale) seems to be acceptable as well. Related to residual plots, when deciding on a link function, a trade-off between easy interpretation and less violation of assumptions (leading to more credible inference) has to be made.

A comment on using the log link is that a positive parameter estimate indicates that an increasing value (or change in factor level for categorical variables) leads to an increase in the mean value of the response, and vice versa. What the problem is in terms of interpretation is the magnitude of a significant effect. A short example for a simple GLM is shown to exemplify this. The linear predictor consists of an intercept and two explanatory variables and the link function is logarithmic:

$$ln(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{11}$$

In $ln(\mu)$ the effects of x_1 does not depend on the value of x_2.
in terms of μ, the model is:

$$\mu = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}$$

In which the magnitude of the effect of x_1 depends on the value of x_2.

Another potential issue that has been noted while fitting the same model with different links is that no interaction effects are significant while using the log link. A possible reason could be that the intuition behind interactions is not the same for different link functions. They can be redundant for the log link. This is argumented for by the following example:

If the true relation between the dependent and explanatory variables is non-linear and let's say it can be described fairly accurate by the relation:

$$\mu = exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2) \tag{12}$$

Then, a second order Taylor series expansion about a point (x_1^0, x_2^0) is:

$$
\begin{aligned}
\mu = f(x_1, x_2) = f(x_1^0, x_2^0) + f_{x_1}^{'}(x_1^0, x_2^0)(x_1 - x_1^0) + f_{x_2}^{'}(x_1^0, x_2^0)(x_2 - x_2^0) + \\
\frac{1}{2} f_{x_1}^{''}(x_1^0, x_2^0)(x_1 - x_1^0)^2 + \frac{1}{2} f_{x_2}^{''}(x_1^0, x_2^0)(x_2 - x_2^0)^2 + \\
f_{x_1 x_2}^{''}(x_1^0, x_2^0)(x_1 - x_1^0)(x_2 - x_2^0) + \mathit{Remainder},
\end{aligned}
$$

in which the right hand side looks similar to a linear predictor. It has an "intercept", "main effects" and "interactions". To simplify this expression and to really exemplify that a Taylor series expansion is similar to how a linear predictor can be constructed, the point of evaluation is now $(x_1^0 = 0, x_2^0 = 0)$ and derivatives evaluated at this point, which are constants, are denoted as β_x. The following expression is obtained:

$$\mu = f(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \mathit{Remainder}$$

This expression includes interaction terms, so in order to construct a good linear predictor for a situation where (12) describes data well and no function is applied to μ, first order interactions should be included.

If instead, a log link is used when the true relation is approximately as in (12), the model becomes linear in terms of the linear predictor, as in (11), and the Taylor series expansion would be

$$ln(\mu) = f(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \mathit{Remainder}$$

since the second order derivatives and mixed derivative are equal to zero. The remainder is also zero, so the approximation is in fact exact (but it is written out to show that a Taylor series expansion is made). Thus, interactions are not relevant to include in a linear predictor if the true relation between the dependent and explanatory variables is as in (12) and a log-link is used.

To summarize, no significant interactions in estimation when using the log link could act as a justification that the true relation between *dioPCB* and the explanatory variables is non-linear, similar to (12).

This example together with the discussion about residuals plots, leads to a decision to base inference on models fitted with a log link. For the inverse and identity link, the residual variance decreases for larger fitted values. Residual plots for log link look better in terms of equal variance, and in the gamma log link model, effects considered as interaction effects can be "hidden" or "implicit". In that case, it is not adequate to specify interaction effects in the model.

5 Results

In this section, results from estimating the model described in (10) are shown and discussed and a subset of the model where insignificant parameters are removed is fitted. Interpretation of estimates and model evaluation is performed. Model evaluation serves as a part in discussing how trustworthy results from model fitting are. Models are fitted using the *glmer* function in the *lme4*-package in R. In order to ensure convergence of parameter estimates in the iterative procedure, *glmer* requires that continuous explanatory variables are scaled to have zero mean and unit standard deviation. When interpreting parameter estimates for these variables, this is taken into consideration. The number of quadrature points used is one. An attempt was made to increase accuracy by increasing the number of points but it was not successful. An increase lead to a singular fit, a consequence of model (10) being large in relation to the number of observations in combination with unbalanced data. Removal of insignificant parameters is done based on an estimation that could have been more accurate. The analysis is made on 224 observations.

In Table 4, results from estimation of model (10) with log-link is displayed.

	estimate & significance
(Intercept)	1.05^{***}
LakeBalticSea	-1.84^{***}
LakeVättern	-0.29
Fat	0.32
binLengthlong	0.19^{**}
Year	-0.11^{**}
CFS	-0.16^{**}
Seasonspring	0.24^{*}
Seasonsummer	0.96^{**}
Seasonwinter	0.35^{*}
LakeBalticSea:Fat	0.31
LakeVättern:Fat	0.17
Fat:binLengthlong	0.06
Fat:Seasonspring	0.17
Fat:Seasonsummer	0.45
Fat:Seasonwinter	0.15
AIC	1187.20
BIC	1248.61
Log Likelihood	-575.60
Num. obs.	224
Num. groups: Location	38
Var: Location (Intercept)	0.13
Var: Residual	0.32

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4: Results from estimation of model (10) with log-link. Hypothesis tests are Wald *t*-tests. Continuous variables are scaled to zero mean and unit standard deviation. Laplace approximation used to approximate the likelihood. Lake=Vänern, Season=Fall, binLength=short are reference categories.

The hypothesis tests reported are Wald *t*-tests. Advantages with these are that they are easily calculated. A downside is that they might not work well for smaller samples sizes, since the test is based on normality but the test statistic is only asymptotically normally distributed. An alternative is to perform likelihood ratio tests. The likelihood ratio test (LR-test) is asymptotically equivalent to a Wald *t*-test but has been shown to perform slightly better for smaller sample sizes (Tuerlinckx et al. 2006). A drawback is that two models have to be fitted in order to test a parameter. LR-tests were performed to test all parameters in model 10 but results are not displayed here, since the LR-tests gave the same results as the Wald *t*-tests.

Insignificant variables are now removed from the model. Regarding the interactions, all include *Fat*, giving some reason to believe that the interactions might hide the main effect, so the main effect is kept in the model. The model without interactions is fitted and results from this estimation is seen in Table 5. It is now clear that removal of the interactions led to *Fat* being significant.

	estimate & significance
(Intercept)	1.07^{***}
LakeBalticSea	-1.79^{***}
LakeVättern	-0.33
Fat	0.54^{***}
binLengthlong	0.20^{**}
Year	-0.11^{**}
CFS	-0.13^{**}
Seasonspring	0.19
Seasonsummer	0.73^{***}
Seasonwinter	0.29
AIC	1180.90
BIC	1221.84
Log Likelihood	-578.45
Num. obs.	224
Num. groups: Location	38
Var: Location (Intercept)	0.14
Var: Residual	0.32

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 5: Results from estimation of model (10) without interactions and with log-link. Hypothesis tests are Wald t-tests. Continuous variables are scaled to zero mean and unit standard deviation. Laplace approximation used to approximate the likelihood. Lake=Vänern, Season=Fall, binLength=short are reference categories.

5.1 Lake effect

The result for differences between bodies of waters seen in Table 5 are separate tests for Lake Vänern and the Baltic Sea compared to Lake Vänern. To account for the family-wise error rate, Tukey's multiple comparison procedure is performed, with results seen in Table 6. The test is complemented with a boxplot of levels for the different waters (Figure 6). Details about Tukey's test are not presented here, but can be found in for example (Montgomery 2017). Results indicate that levels of *dioPCB* is generally lowest for the Baltic Sea. No significant difference was found between Lake Vänern and Lake Vättern. In reality, a difference is believed to exist between Lake Vänern and Lake Vättern, but this difference is likely to be related to generally lower fat percentages in Lake Vättern.

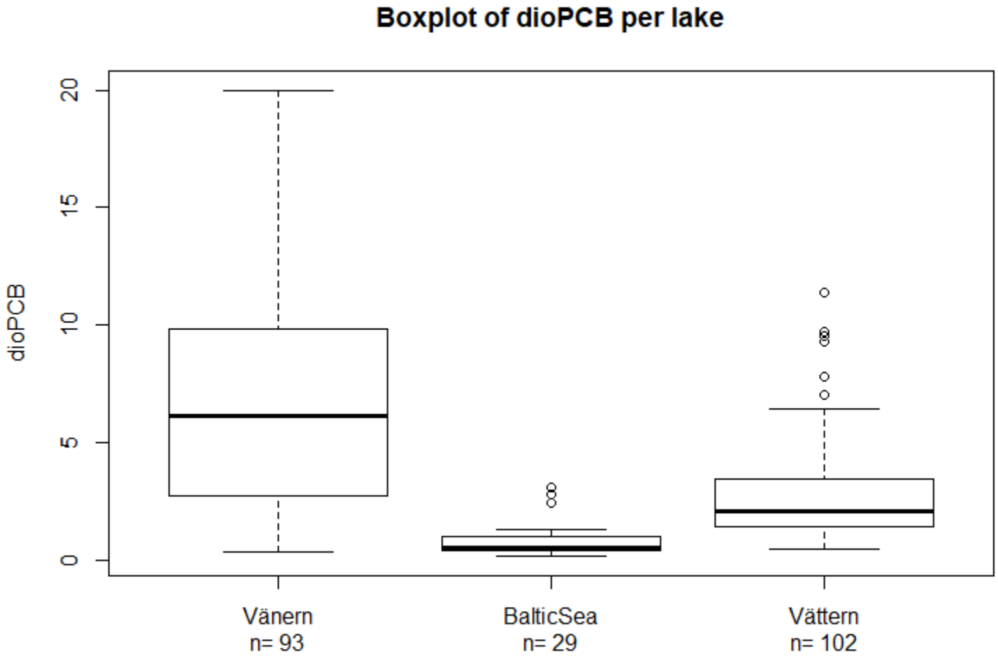


Figure 6: Boxplot of dioPCB in European whitefish per body of water.

H_0	estimate	std.error	$\Pr(>\|z\|)$
BalticSea-Vänern=0	-1.78	0.29	< 0.001
Vättern-Vänern=0	-0.33	0.22	0.31
Vättern-BalticSea=0	1.47	0.30	< 0.001

Table 6: Multiple comparison hypothesis test of Lake-effect using Tukey's procedure. Significance level of 5% is used. Alternative hypothesis is two-sided.

5.2 Morphometry, "Year" and fat content

Regarding fat content, the modelling procedure indicates that higher fat content is related to higher levels of *dioPCB* in European whitefish. Length seems to have an influence as well. Longer fish tend to have higher levels. Year is also significant, and the estimate is negative, indicating that with time, levels of *dioPCB* are reducing. The somatic condition factor is significant, indicating that smaller values of the condition factor is associated with lower levels of *dioPCB*.

Values of the continuous variables have been rescaled. In order to obtain the estimates in terms of the original scale, estimates in Table (5) have to be divided with the original scale standard deviation. Standard deviations of continuous variables in original scales are seen in Table 7.

Variable	st.dev
Fat	2.43
Year	1.01
CFS	0.14

Table 7: Standard deviation of continuous variables in original scale.

Interpretations of magnitudes in terms of the log-link are not so meaningful. What can be done is to estimate how large the relative change is for a unit change of a variable, instead of how large the absolute change is. The relative change is:

$$\text{relative change} = \frac{\hat{\mu}|X_i = x_i}{\hat{\mu}|X_i = x_i + 1} = e^{\hat{\beta}_i}$$

For example, a unit increase in fat content would lead to a relative change in *dioPCB* of $e^{(0.51/2.43)} = 1.23$. Expressed in words: By increasing the fat content with one unit, the level of *dioPCB* is expected to increase by 23%. Expected absolute changes have to be calculated by deciding on values for the fixed parameters.

5.3 Season effect

Results from estimation display differences between spring-fall, summer-fall, and winter-fall. In order to test all possible combinations, and to take the family-wise error rate into account, Tukey's multiple comparison procedure is applied.

First, to get an indication of the differences, a boxplot of *dioPCB* for every season is shown in Figure 7.

Results from Tukey's multiple comparison procedure is seen in Table 8. The only significant difference is between summer and fall, with summer having the lowest levels. Interpretation should be done with caution due to structure of data. As mentioned earlier, no European whitefish was caught during summertime in Lake Vänern, so extrapolating results to apply for Lake Vänern should not be made. In terms of a control fishing program, the formal tests do not indicate that fishing at any specific season or season is preferable. Also, as for the lake effect, there is a belief that differences between seasons with respect to *dioPCB* are due to variations in fat content between the seasons. Inclusion of interactions was the original way to try to investigate this, but they were argumented to be redundant for the log link. The best way to figure out how fat content is related to changes in seasons and different bodies of waters is to construct a model with fat content as the dependent variable, but it is not done in this thesis.

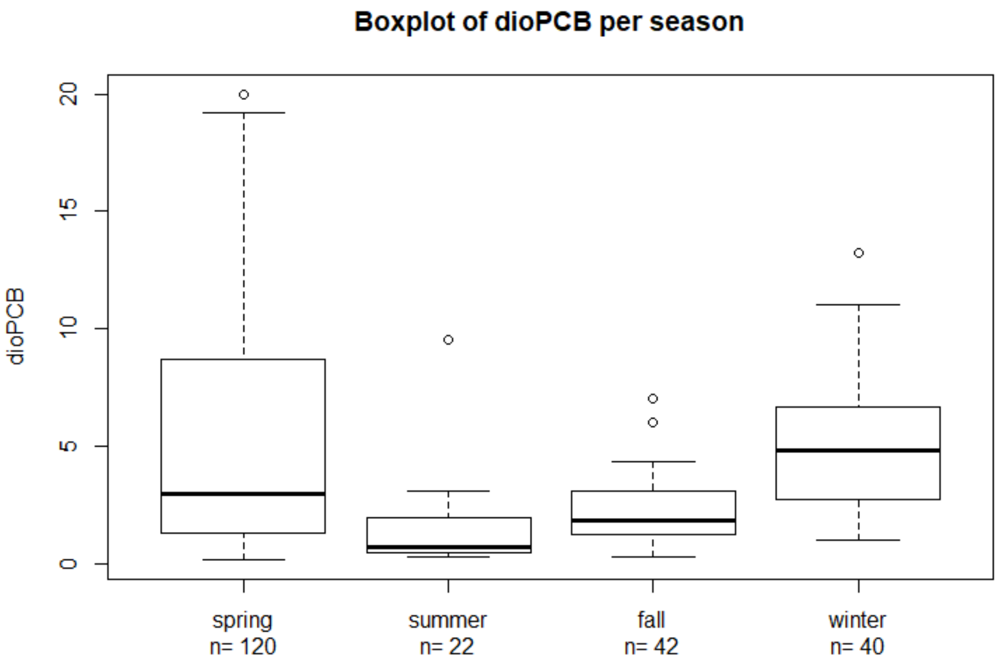


Figure 7: Boxplot of dioxin-like compounds in European whitefish per season.

H_0	estimate	std.error	Pr(> $\|z\|$)
spring-fall=0	0.19	0.11	0.28
summer-fall=0	0.73	0.20	0.0019
winter-fall=0	0.29	0.16	0.21
summer-spring=0	0.54	0.21	0.053
winter-spring=0	0.10	0.16	0.91
winter-summer=0	-0.44	0.25	0.27

Table 8: Multiple comparison hypothesis test of Season-effect using Tukey's procedure. Significance level of 5% is used. Alternative hypothesis is two-sided. Significant difference only between summer and fall.

5.4 Inference about random effect - "Location"

When it comes to the random effect, which in this model is a random intercept corresponding to *Location*, inference is more difficult than for fixed effects. A frequentistic approach is to test the hypothesis $H_0 : \tau^2 = 0$ vs $H_1 : \tau^2 > 0$, where τ^2 is the random effect variance. One problem is that zero is the lowest possible value for the variance component, so the estimate can not "move" freely within the range of the hypothesis test. It is concluded in literature, such as in (Goldman and Whelan 2000) that such tests are conservative. Another problem is that using a likelihood-ratio test might not make sense when only one random effect is in the model, the reason being that the nested model without the random effect is not a generalized linear mixed model but a generalized linear model. Generalized linear mixed models and generalized linear models use different techniques for obtaining maximum likelihood estimates, so comparisons of goodness-of-fit statistics such as AIC might be misleading. Also, it not so clear how many degrees of freedom to assign to the random effect. The random effect often has many levels, so the degrees of freedom could be the number of levels minus one, but in the model only one variance component is estimated, corresponding to one degree of freedom. Still, the authors of *lme4* argues that "With recent versions of *lme4*, goodness-of-fit (deviance) can be compared between (g)lmer (generalized mixed) and (g)lm (generalized) models" (Bolker 2019), so such a comparison is performed here. The model with the random effect is assumed to have one degree of freedom more than the model without. Output seen in Table 9 indicates that *Location* contributes with a significant amount of variability in *dioPCB* levels in European whitefish. This test should be interpreted with caution, as explained by the discussion in this section.

	Df	AIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)
no random effect	11	1236.65	-607.32	1214.65			
with random effect	12	1180.90	-578.45	1156.90	57.75	1	0.0000

Table 9: Likelihood-ratio test of significance of random effect *Location*. Null hypothesis of no random effect variance is rejected.

Leaving formal tests of the random parameter, another perspective of its importance can be made by looking at the estimate and compare it with the residual variability. The estimated *Location* variability is $\hat{\tau}^2 = 0.14$, and the unexplained variability is $\hat{\sigma}^2 = 0.32$. It means that the largest part of the variation in the model is due to unexplained variation, not variation that is a contribution from the catch-sites. This, in combination with the inference made about differences between bodies of waters, indicates that in terms of reducing *dioPCB* in a catch, it is more efficient to change body of water than to change site within a body of water.

Regarding *Location* as a part of the model, there are two choices: to keep it in the model or to remove it. The likelihood-ratio test indicates that *Location* should be kept in the model (although the test should be interpreted cautiously). The choice is to keep it in the model, since the indications to drop it is not very strong. Removing it from the model would result in the model being a generalized linear model, in which the likelihood $L(\mathbf{y}; \boldsymbol{\beta}, \phi)$ does not involve an intractable integral.

5.5 Normality assumption

Here, the normality assumption is evaluated. Normality is checked for both the deviance residuals and the random intercept associated to every level of *Location*. Results are seen in Figure 8 and Figure 9. A slight s-shape is seen in the plots, which might correspond with the potential non-linear pattern seen in residual plot in Figure 3. For the other links, the s-shape was less apparent. Overall, the points lie on a straight line, especially in the middle, which is the most important region. No severe violations of the normality assumptions are seen.

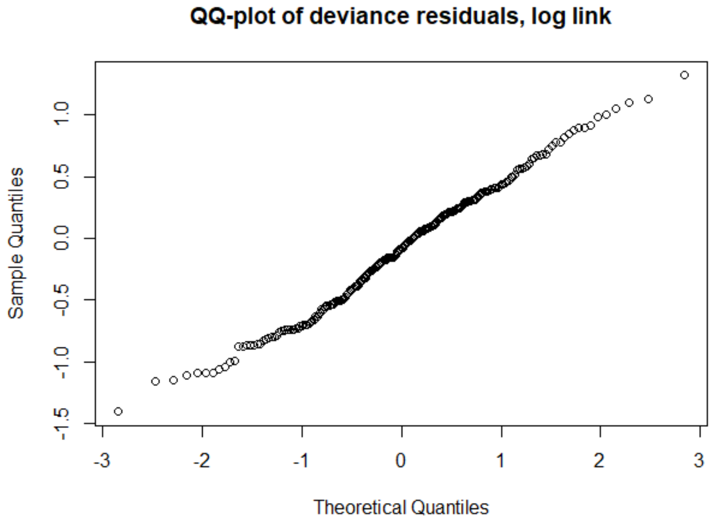


Figure 8: Quantile-Quantile plot of deviance residuals for model 10 fitted with log-link.

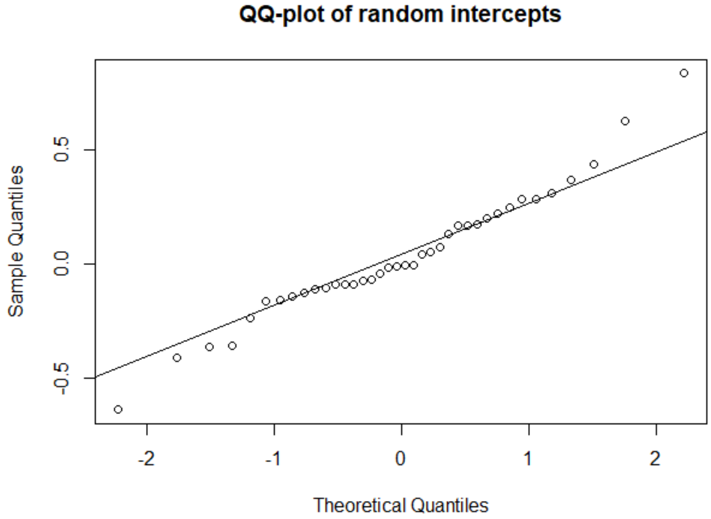


Figure 9: Quantile-Quantile plot of random intercepts model 10 fitted with log-link.

6 Discussion

An attempt was made to, given available data, answer how morphometric, spatial and temporal changes affect DLC content. In order to do this, a generalized linear mixed model was constructed, with arguments about the range and distribution of the response variable *dioPCB* and a discussion about treating explanatory variables as fixed or random. A generalized linear mixed model was chosen, with the catch-site (*Location*) being treated as a random effect and *dioPCB* as gamma distributed. A model was fitted with different link functions and the suitability of each was discussed. A choice was made to base inference on models fitted with a log link. To evaluate the severity of violating assumptions, QQ-plots and plots of residuals vs fitted values were shown and discussed in order to have an idea of how credible inference about parameter estimates are. Assumptions were not violated enough to view inference as poor. In this section, the results from estimation are discusses in terms of previous believes that have emerged during the project ”Dioxins in oily fish - threats and potential for development of small scale coastal and lake fishing” and also in terms of a control fishing program.

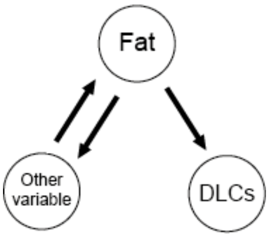
The modelling procedure indicated that morphometric changes lead to changes in *dioPCB*. Fish longer than 38 cm was indicated to in general have higher levels as compared to fish shorter than 38 cm. This is in line with previous believes. Regarding length, a proposal to a possible control fishing program has been to only allow sales of European whitefish shorter than 43 cm. Data did not support this cutoff value as part of an analysis. Too few fish were longer than this value. Instead, 38 cm was used. Whether or not results can be extrapolated to fish longer or shorter than 43 cm is not discussed here, but is left for experts in the field. The somatic condition factor was significant with a negative sign, and the previous belief was that this would have had a positive sign if significant. A high condition factor means that the fish has a high weight in relation to its length. This might be discussed or evaluated further outside this thesis (also, there is a chance of having conducted a type one error).

Seasonal changes were shown to have some influence on the levels of *dioPCB*. Significant differences were found between summer and fall, with summer having lower levels. A visual inspection of the boxplot of *dioPCB* indicated that summer might have lower levels compared to spring and winter as well, but formal tests did not support this. The varying amount of observations per season gave an indication that statistical power would have been better if data were more balanced. The significant difference between summer and fall should not be relied upon in a control fishing program, the reason being that fish caught during summer have only been caught in Lake Vättern and the Baltic sea. Having observations from Lake Vänern as well, where levels of DLC tend to be higher, is likely to result in higher levels for summer, making summer more similar to the other seasons. Also, it would have been preferable if it was known to what sub-species an individual is. A possible season effect comes from an idea that spawning leads to changes in fat content. If, for example, fish spawn in spring, fat content could be lower for this season. The issue is that European whitefish consists of a number of different sub-species, spawning at different seasons, possibly confounding a season effect.

If there is a need to increase the power of formal tests for this effect, more observations would have to be collected for seasons with few observations and having information about sub-species is preferable, but it might be impossible or difficult to find it out. As it appears now, in terms of a control fishing program, there is no strong indication that any season or seasons are better to fish at than others. A previous belief was that there might be a season effect, but results from estimation does not support this belief.

Catch-site (*Location*) was concluded to have an effect on levels of *dioPCB*, although not very large. Further investigation would possibly include an attempt to find catch-sites with generally lower levels of *dioPCB*. Here, emphasis was in obtaining and evaluating an estimate of the variability, not evaluating levels of individual sites, or giving suggestions of possible new sites to fish at. The conclusion that the *Location* effect is not very large is in accordance with with previous believes. In (Hållén and Karlsson 2018) it stated that most likely, the situation regarding DLCs today is mostly affected by atmospheric deposition rather than locally polluted sediment. Also, European whitefish is not a stationary species. Catching one individual at a location does not mean that it has lived its whole life in that area. It could have been feeding in more or less polluted areas, thus ”hiding” a local pollution.

Regarding fat content, a belief that it is strongly related to *dioPCB* was enhanced. In connection to beliefs about other interesting variables related to DLC content, some were believed to be connected to fat content. For example, variations between different bodies of waters was believed to come from generally higher fat content in Lake Vänern, and a season effect could possibly be connected to fat content as well. Put simply, relations could be as:



Such relations are suggested to be examined further, possibly by fitting a model with fat content as dependent variable and *Lake*, *Season* and other variables of interest in the linear predictor. Still, given fat content, the Baltic Sea was shown to have significantly lower levels of *dioPCB* then both Lake Vänern and Lake Vättern.

The amount and structure of data was found to partially limit to what extent it could be analyzed. A lot of explanatory variables and their interactions were interesting, close to occupying to many degrees of freedom needed for a good model fit. A more accurate approximation of the likelihood could have been made by increasing the number of nodes used in the adaptive Gauss-Hermite quadrature, but data did not support this.

When it comes to collecting data, it was not optimally performed from a statistical point of view. In terms of statistical power, effort should have been put on obtaining more balanced data. The number of fish caught at different sites varies greatly, as well as the number of observations per season and body of water. A plus is that the catch-sites are evenly distributed in the waters of interest. Considering financial aspects, it is of interest to get as much information as possible for a given budget. Potentially, a lower number of observations can be collected for the same statistical power, given that the data collecting procedure is constructed with an idea about how to get the most information (given a budget or time restriction). However, implementing this in this project would not be reasonable from a practical perspective. If a fisherman catches ten fish at one spot, and two at another, of course data will be unbalanced and it is not reasonable to tell a fisherman to go to a certain spot at a certain time point and to catch a certain number of fish. It is also the case that fishing is more common for different seasons. Still, the research questions are regarded answered with some confidence, and connected to the discussion about planning studies with regard to statistical power, it is left to the parts involved in the project to discuss what can be done about this in the final year of the project, and in future projects. A final comment is that more fish are continuously being analyzed and added to data throughout 2019, and information about an individuals age will be added, so results and conclusions drawn here can be updated in the future.

7 References

Bolker, Benjamin (2019). *GLMM FAQ.* URL: `https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html` (visited on 04/09/2019).

Breslow, N.E and D.G Clayton (1993). "Approximate inference in generalized linear mixed models". In: *Journal of the American statistical Association* 88.421, pp. 9–25.

Cantillana, T and M Aune (2012). *Dioxin-och PCB-halter i fisk och andra livsmedel 2000-2011.* Livsmedelsverket.

Commission, European (2011). "Commission Regulation (EU) No 1259/2011 of 2 December 2011 amending Regulation (EC) No 1881/2006 as regards maximum levels for dioxins, dioxin-like PCBs and non dioxin-like PCBs in foodstuffs". In: *Off. J. Eur. Union L* 320, pp. 18–23.

Dobson, A.J and A.G Barnett (2018). *An introduction to generalized linear models.* CRC press.

Faraway, J.J (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models.* Chapman and Hall/CRC.

Goldman, N and S Whelan (2000). "Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics". In: *Molecular Biology and Evolution* 17.6, pp. 975–978.

Hadfield, J.D (2010). "MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package". In: *Journal of Statistical Software* 33.2, pp. 1–22.

Hållén, J and M Karlsson (2018). "Dioxiner i sediment från Vänern och Vättern". In: *IVL Swedish Environmental Research Institute* Report number B-2310.

Karlsson, M, G Andersson, P Bohman, J Hållén, A Sandström, and T Viktor (2018). "Dioxiner i fet fisk - Hot och utvecklingsmöjligheter för svenskt småskaligt kust- och insjöfiske". In: *IVL Swedish Environmental Research Institute* Report number B-2301.

Lindstrom, M.J and D.M Bates (1990). "Nonlinear mixed effects models for repeated measures data". In: *Biometrics*, pp. 673–687.

Malmaeus, M and M Karlsson (2014). "Optimerat utnyttjande av lax och strömming från Bottniska viken - förstudie med förslag till provtagningsprogram". In: *IVL Swedish Environmental Research Institute* Report number-B2211.

Malmaeus, M, M Karlsson, and M Rahmberg (2012). "Bottensedimentens roll för dioxinsituationen i industrirecipienter". In: *IVL Swedish Environmental Research Institute* Report number-B2053.

McCullagh, P and J.A Nelder (1989). *Generalized linear models.* 2. ed. London: Chapman & Hall.

McCulloch, C.E (1997). "Maximum likelihood algorithms for generalized linear mixed models". In: *Journal of the American statistical Association* 92.437, pp. 162–170.

Montgomery, D.C (2017). *Design and analysis of experiments.* John wiley & sons.

Nelder, J.A and R.W.M Wedderburn (1972). "Generalized linear models". In: *Journal of the Royal Statistical Society: Series A (General)* 135.3, pp. 370–384.

Nyberg, E, S Danielsson, U Eriksson, S Faxneld, A Miller, and A Bignert (2014). “Spatiotemporal trends of PCBs in the Swedish freshwater environment 1981–2012”. In: *Ambio* 43.1, pp. 45–57.

Rabe-Hesketh, S and A Skrondal (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman and Hall/CRC.

Rabe-Hesketh, S, A Skrondal, and A Pickles (2002). “Reliable estimation of generalized linear mixed models using adaptive quadrature”. In: *The Stata Journal* 2.1, pp. 1–21.

Tuerlinckx, F, F Rijmen, G Verbeke, and P De Boeck (2006). “Statistical inference in generalized linear mixed models: A review”. In: *British Journal of Mathematical and Statistical Psychology* 59.2, pp. 225–255.

Van den Berg, M, L.S Birnbaum, M Denison, M De Vito, W Farland, M Feeley, H Fiedler, H Hakansson, A Hanberg, L Haws, et al. (2006). “The 2005 World Health Organization reevaluation of human and mammalian toxic equivalency factors for dioxins and dioxin-like compounds”. In: *Toxicological sciences* 93.2, pp. 223–241.

8 Appendix

Code from simulation with purpose of verifying that collective samples results in more conservative tests, as compared to having analyzed every individual:

```
# MASS-package for mult.norm sampling
library("MASS")


samples <- 50
r <- 0.2 # correlation
# weights equal to number of values that is taken average of
w <- c(10,rep(1,samples-10))
pvec <- (NA)
pvecm <- (NA)
pvecmw <- (NA)
for (i in 1:1000) {
  #sample data two vectors with correlation
  data <- mvrnorm(n=samples, mu=c(0, 0),
                  Sigma=matrix(c(1, r, r, 1), nrow=2))
  data <- as.data.frame(data)
  #averaging first ten rows to one observation
  meandat <- rbind(colMeans(data[1:10,]),data[11:samples,])
  meandat <- cbind(meandat,w) #adding weights
  htest <- lm(V1~V2, data = data) # fitting regressions
  htestmean <- lm(V1~V2, data = meandat)
  htestmeanw <- lm(V1~V2, data = meandat, weights = w)
  #storing results for tests of "full" dataset
  pvec[i] <- summary(htest)$coefficients[2,4]
  #storing results for tests of when first ten obs are collapsed to mean
  pvecm[i] <- summary(htestmean)$coefficients[2,4]
  #first ten obs collapsed to means and weighted
  pvecmw[i] <- summary(htestmeanw)$coefficients[2,4]
}
#rejection rates in pvec
rrate <- length(pvec[pvec<0.05])/length(pvec)
#rejection rates in pvecm
rratem <- length(pvecm[pvecm<0.05])/length(pvecm)
#rejection rates in pvecmw
rratemw <- length(pvecmw[pvecmw<0.05])/length(pvecmw)
```