TECHNISCHE UNIVERSITÄT WIEN

Generalized linear models with compositional data

DIPLOMARBEIT

angefertigt am

Institut für Stochastik und Wirtschaftsmathematik

Technische Universität Wien

unter der Anleitung von

Ao. Univ. Prof. Dipl. Ing. Dr. techn. Peter Filzmoser

vorgelegt von

Mair Elisabeth

Matrikelnr: 0040372

I-39030 Vintl (BZ), Tulpeweg 9

Wien, 22. Oktober 2015 Elisabeth Mair

Ein herzliches Dankeschön an meine Familie für die ständigen Bestärkungen.
Danke an Herrn Professor Peter Filzmoser für seine Unterstützung und Geduld.

Abstract

In this diploma thesis generalized linear models are adapted to compositional data. Compositional data describe the parts of some whole and carry only relative information. They should not be used directly with generalized linear models as the interpretation of the model can become misleading. The data need to be appropriately transformed. We use the isometric logratio (ilr) transformation proposed by Hron et al. (2010, 2012). In application of this special ilr transformation a meaningful interpretation of the unknown parameters and the inference statistics is possible.
We implemented generalized linear models with compositional data sets in the statistical software R. A model for binary data and a model for count data were adapted. We used a compositional data set resulting from the GEMAS (Geochemical mapping of agricultural and grazing land soils) project to investigate the difference in soil composition in northern and southern Europe. Furthermore we used another compositional data set from the project "Biogeochemical exploration of forests as a base for the long-term landscape exploitation in the Czech Republic" to find a relation between traffic volume and chemical composition of moss and thus model traffic induced pollution in Czech Republic.

Kurzfassung

In dieser Diplomarbeit wurden generalisierte lineare Modelle für Kompositionsdaten angepasst. Kompositionsdaten beschreiben Teile einer Gesamtheit und tragen nur relative Informationen in sich. Sie sollten nicht direkt mit generalisierten linearen Modellen verwendet werden, da damit die Interpretation des Modells zu verfälschten Ergebnissen führen kann. Die Kompositionsdaten müssen vor Anwendung im generalisierten linearen Modell in angemessener Weise transformiert werden. Wir verwenden die ilr (isometric logratio) - Transformation nach Hron et al. (2010, 2012). In Anwendung dieser speziellen Transformation ist eine sinnvolle Interpretation der unbekannten Parameter und der Inferenzstatistik möglich.
Generalisierte lineare Modelle für Kompositionsdaten wurden auch in der Statistiksoftware R implementiert. Jeweils ein praktisches Modell für binäre Daten und für Zähldaten wurde angepasst. Dazu wurden zunächst Kompositionsdaten aus dem GEMAS (Geochemical mapping of agricultural and grazing land soils) - Projekt verwendet, um die chemische Bodenzusammensetzung in Nord- und Südeuropa zu untersuchen. Daten aus einem Biomonitoringprojekt in der Tschechischen Republik (Biogeochemical exploration of forests as a base for the long-term landscape exploitation in the Czech Republic) wurden dazu verwendet, um einen Zusammenhang zwischen dem Verkehrsaufkommen und der chemischen Bodenzusammensetzung zu finden.

Contents

1 Introduction

1.1 Scope

The aim of this thesis is the adaptation of generalized linear models to compositional data. Generalized linear models constitute an extension of classical linear models. Compositional data are quantitative descriptions of the parts of some whole, providing only relative information between their components. A data set out of the GEMAS (Geochemical mapping of agricultural and grazing land soils) project, a cooperation project between EuroGeoSurveys and Eurometaux, and a data set of a biomonitoring project in Czech Republic are used to illustrate the methods. The statistical software R is used in the application.

1.2 Outline of the contents of the thesis

Chapter 2 develops the main concepts of generalized linear models (as defined by Nelder and Wedderburn (1972)). It is about the exponential family of distributions, which includes the Normal, Poisson and binomial distributions. Methods of estimation and of statistical inference for generalized linear models are described.
Chapter 3 outlines the basic concepts and procedures regarding compositional data. The first consistent approach of treatment of such data has been proposed by J. Aitchison in the 1980's (Aitchison, 1986). Since then compositional data analysis has made further progress.
In chapter 4 generalized linear models with compositional data are introduced. We consider generalized linear models where a multivariate (non compositional) response variable is predicted by compositional explanatory variables.
Chapter 5 concerns the practical application with the statistical software R. We focus on models for binary and count data. Compositional data sets representing chemical soil composition are used to study the difference in soil composition in northern and southern Europe and the traffic induced pollution in Czech Republic.

2 Generalized linear models

2.1 Introduction

The term "generalized linear models" (GLMs) was introduced by Nelder and Wedderburn (1972). Generalized linear models constitute an extension of classical linear models. Classical linear models may be summarized in the form:

$$\mathbf{E}(\mathbf{Y}) = \boldsymbol{\mu} \quad \text{where} \quad \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}.$$

The components of the random variable $\mathbf{Y}$ (called *response variable*) are independent normal variables with constant variance σ^2. $E(\mathbf{Y})$ is the expected value of the response variable $\mathbf{Y}$. $\boldsymbol{\beta}$ is the vector of parameters, whose values are usually unknown and have to be estimated from the data. $\mathbf{X}$ is the model matrix, whose elements are constants representing levels of categorical explanatory variables or measured values of continuous explanatory variables.
Thus classical linear models are based on the assumption, that the errors follow a Gaussian or normal distribution with constant variance σ^2 and the mean is a linear function of the explanatory variables.
Generalized linear models represent the following generalization of the classical linear models (Dobson, 2002):

- Response variables can have distributions other than the normal distribution and they may even be categorical rather than continuous.
- The relationship between the response and explanatory variables (called *link*) need not be of the simple linear form.

In the following sections we will describe the theory regarding generalized linear models based mainly on McCullagh and Nelder (1989), Dobson (2002) and Fox (2008).

2.2 Components of generalized linear models

A generalized linear model consists of three components (Dobson, 2002):

1. A random variable Y_i (for the ith of n independently sampled observations), whose components are independent. In the original work of Nelder and Wedderburn (1972) the variable had to be a distribution from the so called exponential family of distributions. Further work extended generalized linear models to multivariate exponential families, to certain non-exponential families and to situations, where the distribution of Y_i is unknown (Fox, 2008),
   In this work the statistical theory for generalized linear models will be revisited on the basis of the exponential families.

2. A set of parameters $\boldsymbol{\beta}$ and explanatory variables $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$, that produce a linear predictor $\boldsymbol{\eta}$ given by
$$\boldsymbol{\eta} = \sum_{j=1}^{N} \mathbf{x}_j \beta_j,$$
   where $\boldsymbol{\beta} = (\beta_1, ..., \beta_N)^T$.

3. A monotone link function $g(\cdot)$ which transforms the expectation of the response variable, $\mu_i = E(Y_i)$, to the linear predictor
$$\eta_i = g(\mu_i),$$
   with $\boldsymbol{\mu} = (\mu_1, ...\mu_n)^T$.

2.2.1 Exponential family of distributions

A single random variable Y has a distribution in the exponential family if its probability function can be written in the form (McCullagh and Nelder, 1989)

$$f_Y(y; \theta, \phi) = exp\left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$$

for some specific functions $a(\cdot), b(\cdot)$ and $c(\cdot)$ and parameters θ and ϕ.

The parameter ϕ is called the dispersion parameter and constant over observations. In some families, ϕ takes on a fixed known value; in other families the parameter has to be estimated from the data. The parameter θ is called the canonical parameter and is a function of the expectation $\mu = E(Y)$. The functions $a(\cdot), b(\cdot)$ and $c(\cdot)$ vary from one distribution out of the exponential family to another and are known functions. For $b(\cdot)$ the first and second derivative with respect to θ have to exist. The function $a(\phi)$ has generally the form

$$a(\phi) = \phi/w,$$

where the parameter w is a known prior weight, usually 1.
The mean and variance of Y are given as:

$$E(Y) = \mu = b'(\theta)$$

$$var(Y) = b''(\theta)a(\phi)$$

$b''(\theta)$ depends on the canonical parameter, and therefore on the mean and is called the *variance function* $V(\mu)$.
Many well-known distributions are a member of the exponential family of distributions, such as the Gaussian (normal), binomial, Poisson, Gamma or inverse-Gaussian families of distributions.
For example, the probability function for the normal distribution is

$$f_Y\left(y;\theta,\phi\right) = \frac{1}{\sqrt{(2\pi\sigma^2)}}\, exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}$$

This can be rewritten as

$$f_Y\left(y;\theta,\phi\right) = exp\left\{(y\mu - \mu^2/2)/\sigma^2 - \frac{1}{2}\left(y^2/\sigma^2 + log(2\pi\sigma^2)\right)\right\}$$

so that $\theta = \mu$, $\phi = \sigma^2$, and

$$a(\phi) = \phi, \quad b(\theta) = \theta^2/2, \quad c(y,\phi) = -\frac{1}{2}\left\{y^2/\sigma^2 + log(2\pi\sigma^2)\right\}.$$

An other example out of the exponential family of distributions is the binomial distribution. The binomial distribution is defined as the discrete probability distribution of the number of successes in a sequence of n independent yes/no experiments in which the probability of success π is the same in all trials. The binomial probability distribution function is

$$f_Y\left(y;\theta,\phi\right) = \binom{n}{y}\pi^y(1-\pi)^{n-y}.$$

Taking logs and collecting terms this can be written as

$$log\ f_Y\left(y;\theta,\phi\right) = y\ log\left(\frac{\pi}{1-\pi}\right) + n\ log(1-\pi) + log\binom{n}{y},$$

which corresponds to the general exponential form

$$log\ f_Y\left(y;\theta,\phi\right) = \left\{ \frac{(y\theta - b(\theta))}{a\left(\phi\right)} + c\left(y,\phi\right) \right\}$$

for the following equivalences:
The canonical parameter is the logit of π

$$\theta = log\left(\frac{\pi}{1-\pi}\right).$$

Solving this term for p and rewriting it, we can see that $log(1-p) = -log(1+e^{\theta})$ and therefore

$$b(\theta) = n\ log(1+e^{\theta}).$$

Finally,

$$c\left(y,\phi\right) = log\binom{n}{y}.$$

We see that $a(\phi) = \phi$ and $\phi = 1$.
Calculating the mean and the variance according to the exponential family of distribution, we get the known results from elementary statistics:

$$E(y) = \mu = b'(\theta) = n\frac{e^{\theta}}{1+e^{\theta}} = n\pi$$

$$var(y) = b''(\theta)a(\phi) = n\frac{e^{\theta}}{(1+e^{\theta})^2} = n\pi(1-\pi).$$

2.2.2 Linear predictor

The linear predictor is a linear function of regressors

$$\eta_i = \sum_{j=1}^{N} x_{ij}\beta_j.$$

The regressors are prespecified functions of the explanatory variables and may therefore include quantitative explanatory variables, transformations of quantitative explanatory variables, dummy regressors and so on.
In classical linear models the linear predictor η and the expected value μ are identical. In generalized linear models η is related to μ by a so called link function.

2.2.3 Link functions

The link function $g(\cdot)$ relates the linear predictor η to the expected value μ.

$$\eta_i = g(\mu_i)$$

Examples of common link functions are the identity, log, reciprocal, logit and probit functions (see Table 2.1).

Table 2.1: Common link functions

Link	$\eta = g(\mu)$
Identity	μ
Log	$ln(\mu)$
Inverse	μ^{-1}
Inverse-square	μ^{-2}
Square-root	$\sqrt{\mu}$
Logit	$ln\frac{\mu}{1-\mu}$
Probit	$\Phi^{-1}(\mu)$
Log-log	$-ln\left[-ln(\mu)\right]$
Complementary log-log	$ln\left[-ln(1-\mu)\right]$

Note: μ is the expected value of the response; η is the linear predictor; $\Phi(\cdot)$is the cumulative distribution function of the standard-normal distribution (Fox, 2008).

Each distribution out of the exponential family has a special link function, for which there exists a sufficient statistic equal in dimension to $\boldsymbol{\beta}$ in the linear predictor $\boldsymbol{\eta} = \sum_{j=1}^{N} \mathbf{x}_j \beta_j$. These link functions associated to a certain family are called canonical links. They occur when $\theta = \mu$ with θ the canonical parameter (McCullagh and Nelder, 1989).
Table 2.2 shows the canonical links, together with the range of variation of the response variable and the variance functions for the commonly used exponential families.
Using the canonical link simplifies the generalized linear models, but other link functions may be used as well. These link functions that can be used vary from family to family. For example, logit, probit, log-log or complementary log-log links are for binomial data. However, it would not be promising to use the identity, log, inverse, inverse-square or square-root links with binomial data (Fox, 2008).

Table 2.2: Canonical link, response range, and conditional variance function for exponential families (Fox, 2008)

Family	Canonical link	Range of Y_i	$var(Y_i/\eta_i)$
Gaussian	Identity	(-∞,+∞)	ϕ
Binomial	Logit	$\frac{0,1,...,n_i}{n_i}$	$\frac{\mu_i(1-\mu_i)}{n_i}$
Poisson	Log	0,1,2,...	μ_i
Gamma	Inverse	(0,∞)	$\phi\mu_i^2$
Inverse-Gaussian	Inverse-square	(0,∞)	$\phi\mu_i^3$

Note: ϕ is the dispersion parameter, η_i is the expectation of Y_i (the response). In the binomial family, n_i is the number of trials.

2.3 Estimation

The principal method of estimation used for all generalized linear models is maximum likelihood. Fitting the models to data by this method, yields estimates of the regression coefficients and also estimated asymptotic standard errors of the coefficients are provided (Fox, 2008).

2.3.1 Maximum likelihood estimation

Let $f(\mathbf{y};\boldsymbol{\theta})$ be the joint probability density function or probability distribution for an observation vector $\mathbf{y} = [y_1, ..., y_n]^T$ given the parameter vector $\boldsymbol{\theta} = [\theta_1, ..., \theta_p]^T$. The likelihood function $L(\boldsymbol{\theta};\mathbf{y})$ is algebraically the same as the probability function, but the probability function is a function of the data with the value of the parameter fixed, while the likelihood function is a function of the parameter with the data fixed.

$$L(\boldsymbol{\theta};\mathbf{y}) = f(\mathbf{y};\boldsymbol{\theta})$$

The maximum likelihood estimator of $\boldsymbol{\theta}$ is the value $\widehat{\boldsymbol{\theta}}$ which maximizes the likelihood function for all $\boldsymbol{\theta}$ in the parameter space Ω.

$$L(\widehat{\boldsymbol{\theta}};\mathbf{y}) \geq L(\boldsymbol{\theta};\mathbf{y}) \quad \text{for all } \boldsymbol{\theta} \text{ in } \Omega$$

$\widehat{\boldsymbol{\theta}}$ is the value that maximizes also the log-likelihood function $l(\boldsymbol{\theta};\mathbf{y}) = log\ L(\boldsymbol{\theta};\mathbf{y})$, since the logarithmic function is monotonic.

$$l(\widehat{\boldsymbol{\theta}};\mathbf{y}) \geq l(\boldsymbol{\theta};\mathbf{y}) \quad \text{for all } \boldsymbol{\theta} \text{ in } \Omega$$

The work with the log-likelihood function is often easier than the work with the likelihood function itself.
The estimator $\widehat{\boldsymbol{\theta}}$ is usually obtained by differentiating the log-likelihood function with respect to each element θ_j of $\boldsymbol{\theta}$ and solving the simultaneous equations

$$\frac{\delta l(\boldsymbol{\theta}; \mathbf{y})}{\delta \theta_j} = 0 \quad \text{for } j = 1, ..., p.$$

To check that the solutions do correspond to maxima it is necessary to verify that the matrix of second derivatives

$$\frac{\delta^2 l(\boldsymbol{\theta}; \mathbf{y})}{\delta \theta_j \delta \theta_k}$$

is negative definite (Dobson, 2002).

In practice however, numerical approximations are used to find $\widehat{\boldsymbol{\theta}}$. McCullagh and Nelder (1989) show, that for generalized linear models the maximum-likelihood estimates of the parameters $\widehat{\boldsymbol{\theta}}$ in the linear predictor η can be obtained by an algorithm based on *iterative weighted least squares*. This algorithm applies to the entire family of distributions for any choice of link function. This is the reason for restricting GLMs to this family (Agresti, 2002). The algorithm McCullagh and Nelder (1989) proposed is equivalent to Fisher scoring and leads to maximum likelihood estimates. Another numerical approximation is the Newton-Raphson method. For the canonical links, where the expected value and the canonical parameter coincide, these two algorithms are equivalent.

2.4 Inference for generalized linear models

2.4.1 Deviance and goodness of fit

McCullagh and Nelder (1989) define the process of fitting a model as replacing a set of data values $\mathbf{y}$ by a set of fitted values $\widehat{\boldsymbol{\mu}}$ derived from a model using usually a relatively small number of parameters. The $\mu's$ equal in general the $y's$ not exactly. The goodness of fit depends on the discrepancy between data values and fitted values. A measure of discrepancy is the so called *deviance*. For n observations, models can be fitted to these containing up to n parameters. In the simplest case the model has only one parameter: a common μ for all the $y's$. This model is called the *null model*.
The model containing n parameters, one per observation is called *full model* or *saturated model*. The $\mu's$ derived from the full model match the data exactly. A full model explains

all variation by the systematic component of the model.
Actually the null model is usually too simple and the full model is not useful. The full model does not summarize the data (and thus not provide data reduction) but repeats them only in full. The full model though is used as baseline for measuring the discrepancy for an intermediate model with p parameters.
The log likelihood can be expressed conveniently in terms of the mean-value parameter $\boldsymbol{\mu}$ rather than the canonical parameter $\boldsymbol{\theta}$ (McCullagh and Nelder, 1989). For $f(y, \theta)$ as density function or probability distribution for the observation y given the parameter θ, then the log likelihood, expressed as a function of the mean-value parameter $\mu = E(Y)$, is

$$l(\mu; y) = log\ f(y; \theta).$$

The log likelihood based on a set of independent observations $y_1, ..., y_n$ is calculated as the sum of the individual distributions:

$$l(\boldsymbol{\mu}; \mathbf{y}) = \sum_i log f_i(y_i; \boldsymbol{\theta}) \quad \text{where} \quad \boldsymbol{\mu} = (\mu_1, ...\mu_n)^T.$$

Let $l(\widehat{\boldsymbol{\mu}}; \mathbf{y})$ for observations $\mathbf{y} = (y_1, ..., y_n)^T$ and means $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^T$ denote the maximum of the log likelihood for the model. For all possible models, the maximum achievable log likelihood is $l(\mathbf{y}; \mathbf{y})$ in a full model.
Now as goodness of fit criterion is used the so called "scaled deviance", which is proportional to twice the difference between the maximum log likelihood achievable and the maximum log likelihood achieved by the model under investigation. The greater the scaled deviance, the poorer the fit:

$$D^*(\mathbf{y}; \boldsymbol{\mu}) = 2l(\mathbf{y}; \mathbf{y}) - 2l(\boldsymbol{\mu}; \mathbf{y}).$$

Table 2.3 summarizes the forms of the deviances for the distributions given in Table 2.2. The deviance is the likelihood-ratio statistic for testing the null hypothesis that the model holds against a more general alternative (i.e. the full model). The deviance function in general though is useful not as absolute measure of goodness of fit but for comparing two nested models (McCullagh and Nelder, 1989; Agresti, 2002). In GLMs for which the dispersion parameter is fixed to 1 (binomial and Poisson GLMs), the likelihood-ratio test statistic is the difference in the deviances for nested models. This difference has under regularity conditions approximate chi-squared null distribution with degrees of freedom equal to the difference between the numbers of parameters in the two models. For GLMs in which there is a dispersion parameter to estimate (Gaussian, gamma, and

Table 2.3: Forms of deviances; summation being over i = 1,...,n (McCullagh and Nelder, 1989)

Gaussian	$\sum(y-\hat{\mu})^2$,
Binomial	$2\sum\{ylog(y/\hat{\mu})-(y-\hat{\mu})\}$,
Poisson	$2\sum\{ylog(y/\hat{\mu})+(m-y)log[(m-y)/(m-\hat{\mu})]\}$,
Gamma	$2\sum\{-log(y/\hat{\mu})+(y-\hat{\mu})/\hat{\mu})\}$,
Inverse-Gaussian	$\sum(y-\hat{\mu})^2/(\hat{\mu}^2 y)$.

inverse-Gaussian GLMs), nested models were compared instead by an incremental F-test (Fox, 2008; Agresti, 2002).

2.4.2 Residuals for GLMs

The dependent variate for normal models can be expressed in the form

$$y = \hat{\mu} + (y - \hat{\mu}),$$

which is datum = fitted value + residual. Residuals can help to explore the adequacy of fit of a model regarding choice of variance function, link function and terms in the linear predictor. Moreover, residuals can give a hint about the presence of anomalous values. For generalized linear models an extended definition of residuals applicable to all the distributions is necessary.
McCullagh and Nelder (1989) define among others the following forms of generalized residuals: Pearson and deviance residuals.

The *Pearson residual* is the raw residual scaled by the estimated standard deviation of Y.

$$r_{i,P} = \frac{y_i - \mu_i}{\sqrt{V(\mu)}}$$

The Pearson residual is the signed square root of the component of the Pearson X^2 goodness of fit statistic, so that

$$\sum_{i=1}^{n} r_P^2 = X^2.$$

The *deviance residual* uses components of the deviance. The deviance is a measure of discrepancy of a generalized linear model. Each unit contributes a quantity d_i to that measure, so that $\sum_{i=1}^{n} d_i = D$. The deviance residual for observation i hence is

$$r_{i,D} = sign(y_i - \mu_i)\sqrt{d_i}.$$

r_D increases with $y_i - \mu_i$ and $\sum_{i=1}^{n} r_{i,D}^2 = D$.

Table 2.4: Choices for the argument family for glm() function in R

Family	**Variance**	**Link**
gaussian	Gaussian	Gaussian
binomial	binomial	logit, probit or cloglog
poisson	Poisson	log, identity or sqrt
gamma	Gamma	inverse, identity or log
inverse.gaussian	inverse Gaussian	1/mu^2
quasi	user-defined	user-defined

2.4.3 Akaike's Information Criterion (AIC)

The Akaike's Information Criterion (AIC) is defined as

$$AIC = -2(\text{maximized log likelihood} - \text{number of parameters in model})$$

This criterion can help to select a good model in terms of estimating quantities of interest. It judges a model by how close its fitted values tend to be to the trues values. The model, that tends to have fit closest to reality is optimal. It is the model with the minimum AIC value (Agresti, 2002).

2.5 Implementation in R

The free and open-source programming language and software environment R provides the `glm()` function to fit generalized linear models, specified by giving a symbolic description of the linear predictor and a description of the error distribution. The function has the following general structure:

```
> glm(formula, family, data, weights, subset, ...)
```

The argument to `glm()` is a model formula with the response on the left of the tilde (read "is modeled as") and a Wilkinson-Rogers model specification formula on the right. The argument family is needed to specify the variance and link function. There are six choices of families available (see Table 2.4). Every family has an associated variance function and a default link function.

3 Compositional data

3.1 Introduction

Compositional data are very common in all experimental fields, especially in environmental and biological sciences. Examples are chemical compositions (chemistry), geochemical elements in geology, body composition in medicine, food composition or soil contamination (environmental sciences). These data are non-negative and sum up to a whole. Hence, no part of this whole can be varied independently from the others. This has wider implications on the treatment of such data. The application of standard statistical methods, that rely mostly on the Euclidean geometry, directly to compositional data can lead to biased and unrealistic results (see, e.g., Filzmoser et al., 2009, 2010). The first consistent approach of treatment of such data has been proposed by J. Aitchison in the 1980's (Aitchison, 1986). This approach is based on the statistical analysis of log-ratios of the original data. Since then in the last 30 years, compositional data analysis has made further progress. The most updated complete work on status and enhancement regarding methods in compositional data analysis is provided by Pawlowsky-Glahn and Buccianti (2011). The analysis of compositional data in this work is reduced to three steps, called principle of working in coordinates (Mateu-Figueras et al., 2011): the representation of data in log-ratio type coordinates, the traditional analysis of these coordinates like real random variables, and the interpretation of the resulting models either in coordinates or expressing the results in terms of the original units. Based mainly on the publication edited by Pawlowsky-Glahn and Buccianti (2011) in the following sections we will describe the basic concepts and procedures regarding compositional data.

3.2 Definition of compositional data

Compositional data (or closed data) in Egozcue and Pawlowsky-Glahn (2011) are defined as data, that quantitatively describe the parts of some whole and provide only relative information between their components. These data are given in form of D-part compositions. A D-part composition is a vector $\mathbf{x} = (x_1, x_2, ..., x_D)$, where the single components - called parts - are positive and carry relative information in form of per-

centages, proportions, concentrations, frequencies or similar. The units of these parts can be percentages, parts per unit, parts per million, molar concentrations or similar. An example for such data is the geochemical composition of rocks. Aitchison (1986) published data sets presenting the mineral compositions of rock specimens. Each composition in this data set consists of the percentages by weight of five minerals (albite, blandite, cornite, daubite and endite). In Table 3.1 an excerpt from a published data set is presented which displays the characteristic features of a compositional data set.

Table 3.1: Excerpt from a data set published in Aitchison (1986). This data set displays the characteristic features of a compositional data set: 1. Each row of the data array corresponds to a single rock specimen, more generally to a composition; 2. each column corresponds to a single mineral, more generally to a part of each composition; 3. each entry is non-negative; 4. the sum of the entries in each row is 1 or equivalently 100 per cent.

Specimen	Percentages by weight of minerals				
no.	albite	blandite	cornite	daubite	endite
H1	48.8	31.7	3.8	6.4	9.3
H2	48.2	23.8	9.0	9.2	9.8
H3	37.0	9.1	34.2	9.5	10.2
H4	50.9	23.8	7.2	10.1	8.0
H5	44.2	38.3	2.9	7.7	6.9

The single parts being components of a whole cannot vary independently of each other. The relevant information is contained in (log)ratios of compositional parts and not in the analytical values themselves. A single part keeps the characteristics of compositional data also if measured alone without other components. This fact is widely neglected as indicated by Filzmoser et al. (2009) in their work.
Compositional data analysis has to fulfill the following principles: *Scale invariance*, *subcompositional coherence* and *permutation invariance*.
Scale invariance means, that scaling the positive components of a vector by a constant leaves equivalent the information carried by the vector. An example would be the change from parts per unit to percentages. Thus vectors with proportional components represent the same composition and form an equivalence class. Such an equivalence class can be represented by a so called closed vector. This is a normalized vector, where the components sum to a given constant κ. This constant can be 1 (in case of proportions), or 100 (in case of percentages) or 10^6 (in case of mg/kg) or any other positive constant. The closure operation, that normalizes any D-part composition $\mathbf{x} = (x_1, x_2, ..., x_D)$ to

the given constant κ (Aitchison, 1986) is defined as:

$$C\mathbf{x} = \left(\frac{\kappa x_1}{\sum_{i=1}^{D} x_i}, \frac{\kappa x_2}{\sum_{i=1}^{D} x_i}, ..., \frac{\kappa x_D}{\sum_{i=1}^{D} x_i} \right).$$

All compositions equivalent to $\mathbf{x}$ and the corresponding equivalence class are represented by the closed vector $C\mathbf{x}$. Similar to the denotation of $\mathbf{x}$, the components of $C\mathbf{x}$ are called parts relative to a total κ.

A subcomposition is defined as a subset of parts of a composition. The analysis of a subcomposition cannot be controversial to the results obtained by an analysis of the full composition. Subcompositional coherence covers the following two criteria:

1. scale invariance is valid for all possible subcompositions;
2. a distance or divergence in the full composition data set shall be greater than or equal to that comparing the corresponding subcompositions (subcompositional dominance).

The second item refers to the fact, that the measurement of distances between compositions and subcompositions should follow the rule of a projection: distances become smaller in a projection.
Permutation invariance means, that the order of the single components has no influence on the results of a compositional analysis. A change in the order produces no different results in the analysis.

3.3 Aitchison geometry

The vector $\mathbf{x} = (x_1, x_2, ..., x_D)$ denotes a D-part composition with sample space the simplex

$$S^D = \left\{ \mathbf{x} = (x_1, x_2, ..., x_D) : x_i > 0 \ (i = 1, 2, ..., D), \sum_{i=1}^{D} x_i = \kappa \right\}.$$

On the simplex Aitchison (1986) introduced the following two operations: perturbation and powering.
Considering the compositions $\mathbf{x}, \mathbf{y} \in S^D$, perturbation of $\mathbf{x}$ with $\mathbf{y}$ is defined as the composition

$$\mathbf{x} \oplus \mathbf{y} = C(x_1 y_1, x_2 y_2, ..., x_D y_D).$$

Powering of $\mathbf{x}$ by a real number α is defined as the composition

$$\alpha \odot \mathbf{x} = C(x_1^\alpha, x_2^\alpha, ..., x_D^\alpha).$$

Both perturbation and powering defined in S^D satisfy the requirements of operations of a vector space and induce a real vector space structure on the simplex.
An inner product on the simplex, with its associated norm and distance are defined as (Pawlowsky-Glahn and Egozcue, 2002; Mateu-Figueras et al., 2011):

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{D} \sum_{i<j} ln\frac{x_i}{x_j}\, ln\frac{y_i}{y_j}\,, \tag{3.1}$$

$$\|\mathbf{x}\|_a = \sqrt{\frac{1}{D} \sum_{i<j} \left(ln\frac{x_i}{x_j} \right)^2}\,, \tag{3.2}$$

$$d_a(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{D} \sum_{i<j} \left(ln\frac{x_i}{x_j} - ln\frac{y_i}{y_j} \right)^2}\,; \tag{3.3}$$

where $\mathbf{x}, \mathbf{y} \in S^D$.
Equipped with this inner product, the vector space on the simplex $(S^D, \oplus, \odot)$ is a finite (D-1)-dimensional Hilbert space. It is a Euclidean vector space structure on the simplex. The algebraic-geometric structure on the simplex as defined above is called Aitchison geometry with Aitchison inner product, Aitchison norm and Aitchison distance. Thus the subindex a is used in Equations (3.1) to (3.3). The Aitchison inner product, norm and distance as well as perturbation and powering satisfy the principles of compositional analysis. Compositional data are represented in the Aitchison geometry on the simplex. Within the Aitchison geometry, $w_1, w_2, ..., w_D$ with $w_i = C(1, 1, ..., e, ..., 1)$ (being e the ith component) for $i = 1, 2, ..., D$ constitute a generating system of S^D. Any vector $\mathbf{x} \in S^D$ can be written as

$$\mathbf{x} = (lnx_1 \odot w_1) \oplus (lnx_2 \odot w_2) \oplus ... \oplus (lnx_D \odot w_D)$$

or equivalently as

$$\mathbf{x} = (ln\frac{x_1}{g_m(\mathbf{x})} \odot w_1) \oplus (ln\frac{x_2}{g_m(\mathbf{x})} \odot w_2) \oplus ... \oplus (ln\frac{x_D}{g_m(\mathbf{x})} \odot w_D), \quad g_m(\mathbf{x}) = \left(\prod_{i=1}^{D} x_i \right)^{1/D}$$

where the coefficients are the centered log-ratio transformation (clr) defined in Aitchison (1986).
From the previous generating system a basis can be obtained by taking any $(D-1)$ vectors such as $w_1, w_2, ..., w_{D-1}$. Then any vector $\mathbf{x} \in S^D$ can be written as

$$\mathbf{x} = (ln\frac{x_1}{x_D} \odot w_1) \oplus (ln\frac{x_2}{x_D} \odot w_2) \oplus ... \oplus (ln\frac{x_{D-1}}{x_D} \odot w_{D-1}).$$

The coefficients correspond to the additive log-ratio transformation (alr) introduced by Aitchison (1986).
The basis $w_1, w_2, ..., w_{D-1}$ is not orthogonal. However, the Euclidean space structure of the simplex allows the construction of an orthonormal basis (or a generating system) and to express the compositions therein. Then all standard statistical methods can be applied to the coordinates and transferred to the simplex.
Such an orthonormal basis can be obtained by applying the usual Gram-Schmidt orthonormalization process to any basis. The constructed basis is not unique. Though the basis should be selected in accordance to the stated problem, which is a not obvious endeavor (Mateu-Figueras et al., 2011).
In S^D as stated in Egozcue and Pawlowsky-Glahn (2011) an orthonormal basis consists of a set of compositions $e_1, e_2, ..., e_{D-1}$ with $\langle e_i, e_j \rangle_a = 0$ for $i \neq j$, and $\|e_i\|_a = 1$. The coordinates of a composition are obtained for a fixed basis as

$$\mathbf{x}^* = ilr(\mathbf{x}) = (\langle \mathbf{x}, e_1 \rangle_a, \langle \mathbf{x}, e_2 \rangle_a, ..., \langle \mathbf{x}, e_{D-1} \rangle_a),$$

with inverse

$$\mathbf{x} = ilr^{-1}(x^*) = \bigoplus_{j=1}^{D-1} x_j^* \odot e_j.$$

This construction of orthonormal coordinates has been called isometric log-ratio transformation (ilr) (Egozcue et al., 2003). The coordinates $x_j^* = ilr_j(\mathbf{x})$ are scale invariant log-ratios (called log-contrasts) and isometric:

$$\begin{aligned} ilr((\alpha \odot \mathbf{x}) \oplus (\beta \odot \mathbf{y})) &= \alpha \cdot ilr(\mathbf{x}) + \beta \cdot ilr(\mathbf{y}), \\ \langle \mathbf{x}, \mathbf{y} \rangle_a &= \langle ilr(\mathbf{x}), ilr(\mathbf{y}) \rangle, \\ \|\mathbf{x}\|_a &= \|ilr(\mathbf{x})\|, \\ d_a(\mathbf{x}, \mathbf{y}) &= d(ilr(\mathbf{x}), ilr(\mathbf{y})). \end{aligned}$$

A particular orthonormal basis $e_1, e_2, ..., e_{D-1}$ is given in Egozcue et al. (2003) with

$$e_i = C\left(\underbrace{exp\left(\frac{1}{\sqrt{i(i+1)}}\right),...,exp\left(\frac{1}{\sqrt{i(i+1)}}\right)}_{i\ elements}, exp\left(-\sqrt{\frac{i}{i+1}}\right), 1, ..., 1\right)$$

and coordinates

$$\mathbf{x} = (y_1 \odot e_1) \oplus (y_2 \odot e_2) \oplus ... \oplus (y_D \odot e_{D-1})$$

with

$$y_i = \frac{1}{\sqrt{i(i+1)}}\, ln\left(\frac{x_1 x_2 \cdots x_i}{(x_{i+1})^i}\right), \quad i = 1, 2, ..., D-1.$$

The coordinates correspond to a particular case of the isometric log-ratio transformation (ilr). The problem with such ilr coordinates (often called balances) is, that they have no interpretation in the sense of the original compositional parts due to the definition of compositional data which contain only relative information. This makes a meaningful interpretation of such variables impossible (Hron et al., 2010).

Another way to construct an orthonormal basis is a method based on sequential binary partitioning of the parts of a composition. With this tool a particular basis in the simplex can be designed, which makes the corresponding coordinates directly interpretable as balances between two groups of parts appearing in some order of the sequential binary partition (Egozcue and Pawlowsky-Glahn, 2005). The Cartesian coordinates of a composition in such a basis are called balances.

Based on sequential binary partition, Hron et al. (2010, 2012) propose the selection of an orthonormal basis, which results in a $(D-1)$-dimensional real vector $z = (z_1, z_2, ..., z_{D-1})$, where the components are defined as

$$ilr(\mathbf{x}) = \mathbf{z} = (z_1, z_2, ..., z_{D-1}).$$

$$z_i = \sqrt{\frac{D-i}{D-i+1}}\, ln \frac{x_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j}} \quad \text{for} \quad i = 1, ..., D-1. \tag{3.4}$$

The inverse transformation of $\mathbf{z}$ to the original composition is given (before closure) by

$$
\begin{aligned}
x_1 &= exp\left(\frac{\sqrt{D-1}}{\sqrt{D}} z_1\right) \\
x_i &= exp\left(-\sum_{j=1}^{i-1} \frac{1}{\sqrt{(D-j+1)\,(D-j)}} z_j + \frac{\sqrt{D-1}}{\sqrt{D-i+1}} z_i\right) \quad \text{for} \quad i = 1, ..., D-1 \\
x_D &= exp\left(-\sum_{j=1}^{D-1} \frac{1}{\sqrt{(D-j+1)(D-j)}} z_j\right)
\end{aligned}
$$

In Equation (3.4) all the relative information of the part x_1 is separated from the remaining parts $x_2, ..., x_D$. The balance z_1 contains all the relative information of part x_1 to all the remaining parts, which are represented by $z_2, ..., z_{D-1}$. z_1 explains all the ratios between x_1 to the other parts of $\mathbf{x}$. The interpretation of z_1 remains unaltered after permutation of the parts $x_2, ..., x_D$ or also when the remaining balances are chosen randomly according to a sequential binary partition of the subcomposition $x_2, ..., x_D$.
Apparently, z_2 does not explain all the relative information about x_2, because the part x_1 is not contained therein. z_2 explains the remaining ratios concerning x_2.
As described in Hron et al. (2012) now an orthonormal basis can be constructed, where the first ilr coordinate explains the compositional part of interest. For it the indices in Equation (3.4) were permuted, so that the compositional part of interest takes the role of x_1. Actually, D different ilr transformations have to be constructed, where the first ilr coordinate explains a different compositional part, respectively.
The D-tuple $(x_1, x_2, ..., x_D)$ in (3.4) is replaced for $l = 1, ..., D$ by

$$(x_l, x_1, ..., x_{l-1}, x_{l+1}, ..., x_D) := (x_1^{(l)}, x_2^{(l)}, ..., x_l^{(l)}, x_{l+1}^{(l)}, ..., x_D^{(l)}).$$

The resulting ilr transformation is

$$z_i^{(l)} = \sqrt{\frac{D-i}{D-i+1}}\, ln \frac{x_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j^{(l)}}} \quad \text{for} \quad i = 1, ..., D-1. \tag{3.5}$$

We have $z_i^{(1)} = z_i$ for $i = 1, ..., D-1$. The above constructed different ilr transformations (see Equation (3.5)) result as expressions of $\mathbf{x}$ in different orthonormal bases on S^D; they are orthogonal transformations of each other (Egozcue et al., 2003).

3.4 Inference in models with compositional explanatory variables

Modeling with compositional explanatory variables can be misleading if the original untransformed data are directly used in the model, because compositional data are not represented in the Euclidean space.
Hron et al. (2012) proposed the following approach in a study about linear regression with compositional explanatory variables. They selected the ilr basis like described above (see Equation (3.4)), where the first ilr coordinate contains all the relative information about one particular compositional part. Thus parameter estimation and also inference statistics for this parameter referred completely to this part. No hypotheses for the parameters referring to the other ilr variables were tested, as the corresponding predictor ilr variables are not assigned to one single compositional part and therefore not immediately interpretable. These variables were only used for regression, but not for interpretation.
Then another ilr basis was chosen, where again the first ilr coordinate contains all information about another specific part. This was done for every explanatory variable.
The fit of every model was exactly the same, as the different ilr transformations are orthogonal rotations of the corresponding bases.

3.5 Treatment of zeros

The application of log-ratio transformations to model compositional data demands no zeros in the data matrix, a requirement of both ratios and logarithms. In compositional data analysis there can be distinguished three different types of zeros: rounded zeros, count zeros and essential zeros.
Rounded zeros appear mostly when dealing with continuous data (e.g. weights, time, length) and are not real zeros. They result from the rounding to zero of very small observed data (called "rounded zeros"). The observed value lays below a particular maximum possible rounding-off error ($\epsilon = 10^d$). Martín-Fernández et al. (2011) assign to this category of zeros also the zeros resulting from very small values, that cannot be recorded. These values are called below-detection values. Both zeros, the rounded zeros and the below detection values are treated in the same way.
Another category of zeros are the count zeros. A count represents a number of times an event occurs. A vector of counts represents categorical data, where every count expresses the number of items falling into each category. A zero value in one category means no observation. This missing observation can be caused by limited size of the sample and

is in this case called a count zero.
Essential zeros, however, mean components, which are real zeros.
Martín-Fernández et al. (2011) summarized the recent techniques of the treatment of zeros in compositional data analysis.

3.6 Implementation in R

Within the free and open-source programming language and software environment R the package RobComposition provides tools for compositional data analysis. The package contains methods for imputation of compositional data including robust methods, (robust) outlier detection for compositional data, (robust) principal component analysis for compositional data, (robust) factor analysis for compositional data, (robust) discriminant analysis (Fisher rule) and (robust) Anderson-Darling normality tests for compositional data as well as popular log-ratio transformations (alr, clr, ilr, and their inverse transformations) (Templ et al., 2015).
The ilr transformation is implemented with the special choice of the balances according to Hron et al. (2010). The function in R is called `isomLR()`. The `isomLR()` transformation moves D-part compositional data from the simplex into a (D-1)-dimensional real space isometrically. From this choice of the balances, all the relative information of the part x_1 from the remaining parts is separated. The input of the function consists of an object of the classes data frame or matrix with positive entries. The function provides the ilr transformed data (Templ et al., 2015).

4 Generalized linear models with compositional data

In this work we consider generalized linear models with compositional explanatory variables. The aim is to estimate parameters from a generalized linear model, when a multivariate (non compositional) response variable is predicted by compositional explanatory variables.

4.1 Generalized linear models with compositional explanatory variables

Generalized linear models are a generalization of the general linear model. They differ from the general linear model in two principal aspects: The response variable can be non-normal distributed and does not have to be continuous; the response variable is connected to a linear combination of explanatory variables (the so called *linear predictor*) via a so called link function, which has not to be linear.
Generalized linear models in real space are defined in terms of a set of independent random variables $Y_1, ..., Y_N$ from a distribution in the exponential family. The explanatory variables $\mathbf{x}_1, ..., \mathbf{x}_N$ are related to a vector $\eta_1, ..., \eta_N$ through a linear model. This linear combination of explanatory variables is the *linear predictor* $\boldsymbol{\eta} = \sum_{j=1}^{N} \mathbf{x}_j \boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$, $p < N$. For model specification a smaller set of parameters $\beta_1, ..., \beta_p$ (where $p < N$) is of interest.
Let the expectation of the response variable be $E(Y_i) = \mu_i$, $i = 1, ..., N$. The model links then μ_i to η_i by $\eta_i = g(\mu_i)$, where the link function g is a monotone differentiable function. Thus, g links $E(Y_i)$ to the explanatory variables through the formula

$$g(\mu_i) = \sum_{j=1}^{p} x_{ij}\beta_j \quad \text{for} \quad i = 1, ..., N.$$

The unknown parameters $\beta_1, ..., \beta_p$ need to be estimated.
In summary, a GLM is a linear model for a transformed mean of a response variable that has distribution in the natural exponential family (Agresti, 2002).
Generalized linear models though are based on the usual Euclidean geometry and on data that carry absolute information. The application with compositional data, where the components of the explanatory variables $\mathbf{x}_i$ carry only relative information, is not

reasonable. Compositional data do not plot into Euclidean space because they have their own geometry which is not linear but curved in the Euclidean sense (Filzmoser et al., 2010). They are represented in the Aitchison geometry on the simplex.
As a solution, the ilr (isometric logratio) transformation can be applied on the compositional explanatory data. Such a transformation expresses the original compositions in $D-1$ orthonormal coordinates with respect to the Aitchison geometry. The construction of the orthonormal coordinates is crucial for the interpretability of the results. Hron et al. (2012) propose an orthonormal basis, where the first ilr coordinate explains the compositional part of interest. A detailed explanation of the equations can be found in Section 3.3.

For a sample of n independent observations of the response variable $Y_i \in \Re$, $i = 1, ..., n$, and a set of explanatory variables $\mathbf{x}_i \in S^D$, $i = 1, ..., n$, a generalized linear model of Y_i, $i = 1, ..., n$, on the ilr transformed explanatory variables $\mathbf{z}_i \in R^{D-1}$ of Equation (3.4) with the unknown parameters b_j, $j = 0, ..., D-1$, can be described as follows:

$$g(\mu_i) = b_0 + \sum_{j=1}^{D-1} z_{ij} b_j \quad \text{for} \quad E(Y_i) = \mu_i, \quad i = 1, .., n.$$

The parameters b_j, $j = 0, ..., D-1$, can be estimated by the maximum likelihood method. b_0, the intercept term, is independent of the choice of the orthonormal basis, as it is related only to the response variable Y_i. The other parameters $b_1, ..., b_{D-1}$ are directly connected to the ilr coordinates.
As Hron et al. (2012) described for the linear regression model with compositional data, we can consider the lth ilr-basis according to Equation (3.5), which leads to the following generalized linear model:

$$g(\boldsymbol{\mu}) = b_0 + b_1^{(l)} z_1^{(l)} + ... + b_{D-1}^{(l)} z_{D-1}^{(l)}$$

$z_1^{(l)}$ explains all the relative information about the compositional part $x_1^{(l)}$, and therefore the coefficient $b_1^{(l)}$ can be assigned to this part. The interpretation of the other parameters $b_j^{(l)}$, $j = 2, ..., D-1$, is difficult, since they are directly connected to all ilr coordinates. Altogether, D different generalized linear models have to be considered. The resulting generalized linear model is

$$g(\mu_i) = b_0 + b_1 z_{i1} + ... + b_{D-1} z_{i,D-1} + e_i \quad \text{for} \quad i = 1, .., n.$$

e_i stands for the error variability that cannot be accounted for by the predictors. The explanatory variables $\mathbf{z}_i = (1, z_{i1}, ..., z_{i,D-1})'$ represent the ilr transformation of $\mathbf{x}_i$ (in-

cluding 1 for the intercept).
The intercept term b_0 and the model fit remain unchanged, since the different ilr bases are orthogonal (Hron et al., 2012).

5 Application with R

In this chapter, generalized linear models are practically implemented for compositional data sets. Models for binary data (binomial distribution) and for count data (Poisson distribution) are studied. We use a compositional data set resulting from the GEMAS (Geochemical mapping of agricultural and grazing land soils) project, a cooperation project between EuroGeoSurveys and Eurometaux, to implement a model for binary data. A compositional data set resulting from a biomonitoring campaign in Czech Republic is used to implement a model for count data.

5.1 Binary variables, binomial regression and logistic model

In this section we apply generalized linear models with binomial random component and a compositional explanatory data set.

5.1.1 The GEMAS Project

GEMAS is an acronym for *Geochemical Mapping of Agricultural and Grazing land Soil* and identifies a cooperation project between the EuroGeoSurveys Geochemistry Expert Group and Eurometaux.
The EuroGeoSurveys Geochemistry Expert Group is a not-for-profit organization representing 33 national Geological Surveys and some regional Surveys in Europe. EuroGeoSurveys provides the European institutions with advice and information in areas such as (EuroGeoSurveys, 2015):

- the use and the management of on- and off-shore natural resources;
- the identification of natural hazards of geological origin, their monitoring and the mitigation of their impacts;
- environmental management, waste management and disposal; land-use planning;
- sustainable urban development and safe construction;
- e- government and the access to geoscientific meta data and data;

- the development of interoperable and harmonized geoscientific data at the European scale.

Eurometaux is the European Association of Metals, the Brussels-based association servicing and representing the European non-ferrous metals industry.

The GEMAS project started in 2008 and ended with the launch of the GEMAS Geochemical Atlas in December 2013. The project provides high quality harmonized, freely and interoperable geochemical data for ploughed agricultural soil and for non-cultivated grass land (non-cultivated for at least 10 years).
Within the project more than 4.000 samples of agricultural soil (0 - 20 cm) and grazing land soil (0 - 10 cm) were collected at an average sample density of 1 site per 2.500 km^2 across 33 European countries, following strictly a field manual. The concentration of almost 60 chemical elements, and the parameters determining their availability and binding in agricultural and grazing soils at the scale of a continent (Europe - 5.6 million km^2 were sampled) were documented (GEMAS, 2015). The data sets were harmonized with respect to

1. land-use (agricultural soil, 0 - 20 cm and grazing land soil, 0 - 10 cm);
2. spatial scale (homogeneous sampling density: 1 site/2.500 km^2 (grid of 50 x 50 km);
3. sample preparation (<2 mm grain size);
4. analytical methodology: Aqua regia extractable (ICP-MS 53 elements), total (XRF, 41 elements) and mobile metal ion (MMI®, 55 elements) concentrations, lead isotope ratios, pH (0.01M $CaCl_2$), Total Organic Carbon, Total Carbon, Total Sulphur, Effective Cation Exchange Capacity (eCEC at pH of the soil, silver thiurea method), mid-infra red (MIR) spectra, Texture (sand, silt, clay) and Partitioning coefficients (k_D-values) for selected elements.

For more information on data, methodology and results of the GEMAS project, in April 2014 was released a set of two volumes, accompanied by a DVD with all the analytical data, maps, diagrams and tables:

- Reimann, C., Birke, M., Demetriades, A., Filzmoser, P. & O'Connor, P. (Editors), 2014. Chemistry of Europe's agricultural soils - Part A: Methodology and interpretation of the GEMAS data set. Geologisches Jahrbuch (Reihe B 102), Schweizerbarth, Hannover, 528 pp. + DVD

- Reimann, C., Birke, M., Demetriades, A., Filzmoser, P. & O'Connor, P. (Editors), 2014. Chemistry of Europe's agricultural soils - Part B: General background information and further analysis of the GEMAS data set. Geologisches Jahrbuch (Reihe B 103), Schweizerbarth, Hannover, 352 pp.

5.1.2 Data set

The GEMAS data set used in this work consists of 2108 observation of 144 variables. Only samples of agricultural soils are included in this data set. The elimination of all incomplete observations leads to a data set with 2056 observations. The following list contains all the variables names.

```
names
  [1] "ID"                  "COUNTRY"             "C_ID"
  [4] "TYPE"                "TYPE2"               "XCOO"
  [7] "YCOO"                "XLAEA"               "YLAEA"
 [10] "ALT"                 "CIA"                 "sand"
 [13] "silt"                "clay"                "sand_norm"
 [16] "silt_norm"           "clay_norm"           "soiltype"
 [19] "soilclass"           "climate"             "MeanTemp"
 [22] "AnnPrec"             "PM"                  "CEC"
 [25] "pH_CaCl2"            "TOC"                 "C_totF"
 [28] "C_tot"               "N_totF"              "S_totF"
 [31] "S_tot"               "CNratio"             "Ag"
 [34] "Al"                  "As"                  "Au"
 [37] "B"                   "Ba"                  "Be"
 [40] "Bi"                  "Ca"                  "Cd"
 [43] "Ce"                  "Co"                  "Cr"
 [46] "Cs"                  "Cu"                  "Fe"
 [49] "Ga"                  "Ge"                  "Hf"
 [52] "Hg"                  "In"                  "K"
 [55] "La"                  "Li"                  "Mg"
 [58] "Mn"                  "Mo"                  "Na"
 [61] "Nb"                  "Ni"                  "P"
 [64] "Pb"                  "Pd"                  "Pt"
 [67] "Rb"                  "Re"                  "S"
 [70] "Sb"                  "Sc"                  "Se"
 [73] "Sn"                  "Sr"                  "Ta"
```

```
 [76] "Te"                "Th"                "Ti"
 [79] "Tl"                "U"                 "V"
 [82] "W"                 "Y"                 "Zn"
 [85] "Zr"                "C_tot.1"           "S_tot.1"
 [88] "SiO2"              "Si_XRF"            "TiO2"
 [91] "Ti_XRF"            "Al2O3"             "Al_XRF"
 [94] "Fe2O3"             "Fe_XRF"            "MnO"
 [97] "Mn_XRF"            "MgO"               "Mg_XRF"
[100] "CaO"               "Ca_XRF"            "Na2O"
[103] "Na_XRF"            "K2O"               "K_XRF"
[106] "P2O5"              "P_XRF"             "LOI"
[109] "As_XRF"            "Ba_XRF"            "Bi_XRF"
[112] "Ce_XRF"            "Co_XRF"            "Cr_XRF"
[115] "Cs_XRF"            "Cu_XRF"            "Ga_XRF"
[118] "Hf_XRF"            "La_XRF"            "Mo_XRF"
[121] "Nb_XRF"            "Ni_XRF"            "Pb_XRF"
[124] "Rb_XRF"            "Sb_XRF"            "Sc_XRF"
[127] "Sn_XRF"            "Sr_XRF"            "Ta_XRF"
[130] "Th_XRF"            "U_XRF"             "V_XRF"
[133] "W_XRF"             "Y_XRF"             "Zn_XRF"
[136] "Zr_XRF"            "X208_207"          "X207_208"
[139] "X208_206"          "X206_208"          "X206_207"
[142] "X207_206"          "SUSCEPTIBILITY"    "SUSCEPTIBILITY.FE2O3"
```

5.1.3 Difference in soil composition in northern and southern Europe

In the GEMAS project large differences in the concentration of many chemical elements between the soil of northern and southern Europe could be observed. In young soils from northern Europe the concentrations for many elements is 2-3 times lower than in older and more weathered southern European soil. Figure 5.1 shows the arsenic (As) map, produced in the GEMAS project. Arsenic concentrations in agricultural soil are clearly higher in southern and western compared to northern Europe.

Given a compositional data set of chemical elements and a binary variable representing the geographical position north or south, we try to fit a generalized linear model that describes the observation made in the GEMAS project.

The compositional explanatory data set used in this work consists of the total concentration of 10 major elements of a soil sample plus Loss on Ignition (LOI). The total

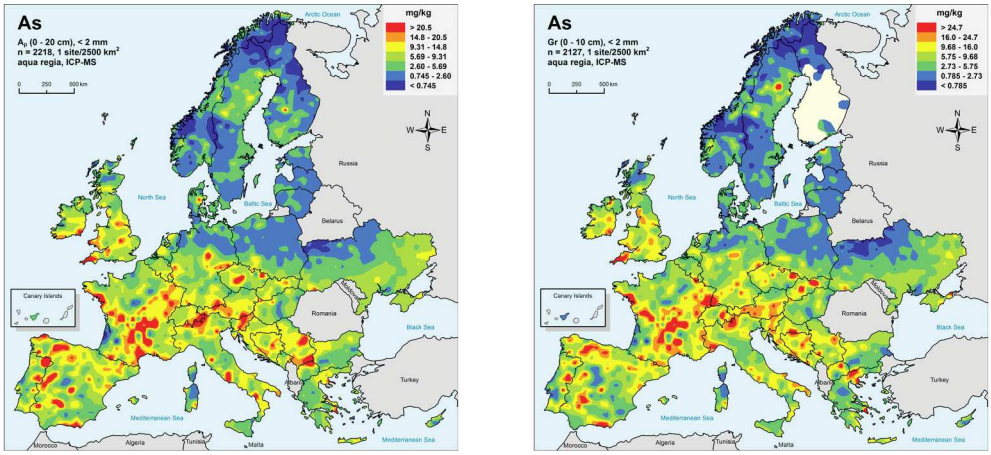

Figure 5.1: Distribution maps for arsenic (As), in agricultural soils (left) and grazing landes (right) (GEMAS, 2015).

concentrations were determined by wavelength dispersive X-ray fluorescence spectrometry (WD-XRFS) using PAN2400 and AXIOS WD-SRFs with Cr- and Rh-anode X-ray tubes, respectively (Reimann et al., 2012). The 10 major elements are Al_2O_3, CaO, Fe_2O_3, K_2O, MgO, MnO, Na_2O, P_2O_5, SiO_2 and TiO_2. Loss on Ignition (LOI) in the GEMAS project was determined on all samples via slowly heating to 1030°C, keeping them at this temperature for 15 min in a muffle furnace, letting them cool to room temperature in a desiccator and reporting the weight loss. As cited in Reimann et al. (2012) this set of total concentrations of the 10 major elements plus LOI is a "classical" example of a "closed" data set.
The respective variables of the data set are the following:

`Al_XRF, Ca_XRF, Fe_XRF, K_XRF, Mg_XRF, Mn_XRF, Na_XRF, P_XRF, Si_XRF, Ti_XRF and LOI.`

As mentioned above, 2056 samples are provided with the data set.
The binary response outcome for each sample is the geographical classification "north" and "south". To create this geographical classification the variable "climate" was used, which consists of the four climate classes Spbo, BoTe, Temp and Medi. The climate classes Spbo and BoTe were grouped into the category "north" and Temp and Medi into the category "south". Figure 5.2 depicts the climate classification for the single sample sites.

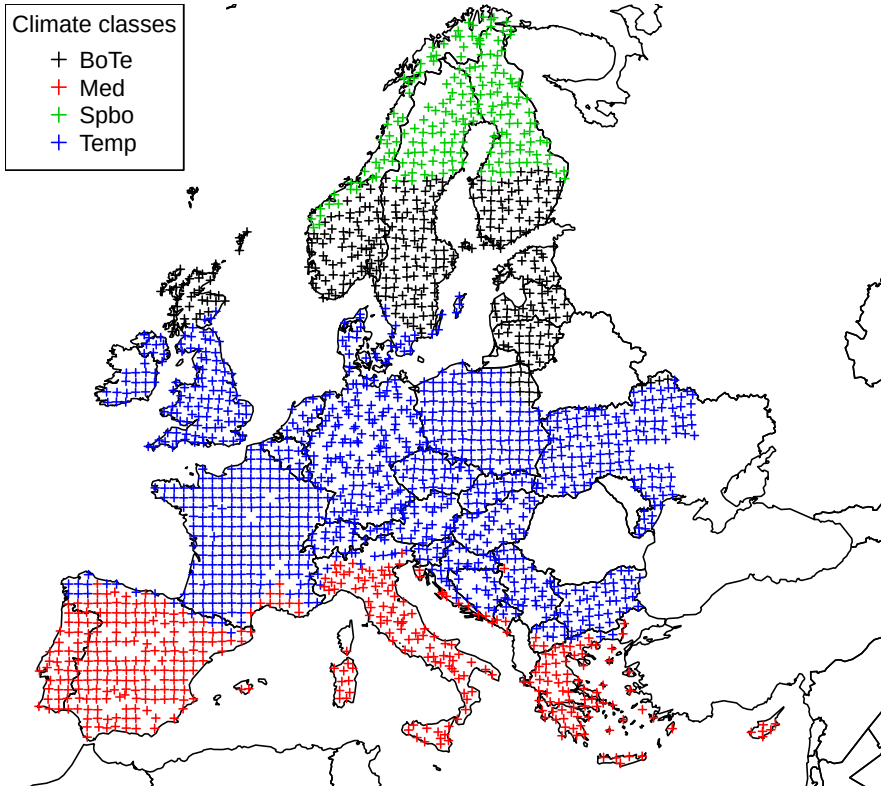

Figure 5.2: Sample sites in the GEMAS study. The single points are colored according to the climate classification.

Thus, the input data is given by a matrix `X1`, representing 2056 compositions of 10 chemical elements, the variable `LOI` and the vector `north`, consisting in a logical operator with TRUE/FALSE elements. If R is forced to do arithmetic on logical values, then it takes TRUE to be 1 and FALSE to be 0.

```
X1 = (Al_XRF, Ca_XRF, Fe_XRF, K_XRF, Mg_XRF, Mn_XRF, Na_XRF, P_XRF,
      Si_XRF, Ti_XRF, LOI)
```

```
north = (FALSE FALSE TRUE TRUE TRUE FALSE ...)
```

Goal is to obtain a model that describes the relationship between north-south geographical position and the ilr-transformed variables `AL_XRF, ..., LOI`.

```
north ~ Al_XRF + Ca_XRF + Fe_XRF + ... + Zn_XRF + Zr_XRF + LOI
```

It is important to note, that the ilr transformed variables are ratios and therefore dimensionless numbers, which have not an obvious meaning to a geochemist studying the concentration of the chemical elements. Reimann et al. (2012) though show, that elements with high concentrations have high ilr variables.

The implementation in R is based on the functions `lmCoDaX()` and `glm()`, already implemented in R. The function `lmCoDaX()`, available in the R-package robCompositions, delivers appropriate inference for regression of y on a compositional matrix X. Classical and robust regression can be applied with compositional explanatory variables. The approach is based on the isometric logratio (ilr) transformation with a special choice of the balances according to Hron et al. (2010).
The function `glm()` is used to fit generalized linear models, specified by giving a symbolic description of the linear predictor and a description of the error distribution.
The code of the function `lmCoDaX()` was adapted to the use with generalized linear models and the function `glm()`. The respective R-Code of this new function called `glmCoDaX()` can be found in the appendix of this work.
The R input for a binomial GLM with logit link function and our data is given by:

```
glmCoDaX(X1,north,binomial)
```

At first we split the compositional data set into two parts, a training set to fit a model and a test set to assess the generalization error for the chosen model. For the training set, randomly 2/3 of all observations are selected; this corresponds to 1370 observations.

Consequently the test set comprehends the remaining 1/3 of all observations, which corresponds to 686 observations. With the compositional training set we try to fit three models. In the first model we use the ilr transformed compositional data set according to Hron et al. (2012). In the second model we use the original compositional data set. Due to the skewness of the data though, we apply a log transformation on them. And finally in the third model we fit clr transformed compositional data.

In the case of ilr transformed variables we get the following result. For the exact structure of the input data see the R-code in the appendix.

```
glmCoDaX

Call:
glm(formula = y ~ ., family = family, data = d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9093  -0.4347  -0.1156   0.1615   2.9469

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  20.8193     3.0824   6.754 1.44e-11 ***
X.Al_XRF    -11.1659     1.0285 -10.856  < 2e-16 ***
X.Ca_XRF     -0.3374     0.1704  -1.980   0.0477 *
X.Fe_XRF      3.1446     0.5114   6.149 7.80e-10 ***
X.K_XRF       4.0287     0.4768   8.449  < 2e-16 ***
X.Mg_XRF      0.4614     0.2911   1.585   0.1129
X.Mn_XRF     -1.3963     0.2739  -5.097 3.45e-07 ***
X.Na_XRF      5.7517     0.3978  14.459  < 2e-16 ***
X.P_XRF      -0.2094     0.2572  -0.814   0.4156
X.Si_XRF     -1.2826     0.2674  -4.797 1.61e-06 ***
X.Ti_XRF     -0.2293     0.3901  -0.588   0.5567
X.LOI         1.2345     0.2693   4.585 4.54e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 1559.19  on 1369  degrees of freedom
Residual deviance:  775.85  on 1359  degrees of freedom
AIC: 797.85

Number of Fisher Scoring iterations: 7
```

Seven iterations were required to fit this model. R flags significant coefficients with one, two or three stars depending on their p-values. We can see that some of the ilr transformed variables are highly significant. These variables are `Al_XRF`, `Ca_XRF`, `Fe_XRF`, `K_XRF`, `Mn_XRF`, `Na_XRF`, `Si_XRF` and `LOI`. It is important to note, that `Na_XRF` for example has a very positive effect on the response variable, `Al_XRF` a negative effect. There is also an intercept term, which is very positive and highly significant.
R returns two forms of deviances, the null deviance and the residual deviance. The deviance is a measure of goodness of fit of a GLM. Higher numbers indicate a worse fit. The "null deviance" is a measure for how well the response variable is predicted by a model that includes only the intercept. All other terms are excluded. Since the null deviance is 1559.19 on 1369 degrees of freedom, the null model is a poor fit. The degrees of freedom for this model are the number of data points n minus 1 if an intercept is fitted. Including the 10 independent variables decreased the deviance to 775.85 points on 1359 degrees of freedom. The residual deviance refers to the fitted model, which has $n - p$ degrees of freedom, where n is the number of data points and p is the number of parameters, including any intercept.
The AIC-value for this model is 797.85. This number though is only meaningful when comparing different models (not necessary nested).
This approach uses 10 different ilr transformations according to the number of explanatory variables. For every ilr transformation the first ilr coordinate contains all information about another specific part. The model fit in this approach however is independent of the ilr transformation selection, since the different ilr transformations are orthogonal rotations of the corresponding bases.

To compare the results, a logistic regression with generalized linear models has been carried out for the original data (log transformed).
In this case the R input is given by:

```
glmfit<-glm(y~.,data=data.frame(y=north,X=log(X1)),family=binomial)
```

R returns the following warning message:

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

This warning means that the data is possibly linearly separable.

A detailed information on the fit can be obtained with the `summary()` function:

```
summary(glmfit)
```

The following result can be obtained:

```
Call:
glm(formula = y ~ ., family = binomial, data = data.frame(y =
    north_train, X = log(X1_train)))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8113  -0.4287  -0.1166   0.1555   2.9417

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   4.37676   12.08901   0.362   0.7173
X.Al_XRF    -10.46106    0.98703 -10.598  < 2e-16 ***
X.Ca_XRF     -0.29372    0.16344  -1.797   0.0723 .
X.Fe_XRF      2.97309    0.49009   6.066 1.31e-09 ***
X.K_XRF       3.77559    0.45604   8.279  < 2e-16 ***
X.Mg_XRF      0.46269    0.27889   1.659   0.0971 .
X.Mn_XRF     -1.34103    0.26165  -5.125 2.97e-07 ***
X.Na_XRF      5.47714    0.37953  14.431  < 2e-16 ***
X.P_XRF      -0.16028    0.24704  -0.649   0.5165
X.Si_XRF     -0.09774    0.84183  -0.116   0.9076
X.Ti_XRF     -0.23100    0.37350  -0.618   0.5363
X.LOI         1.46626    0.33052   4.436 9.16e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 1559.19  on 1369  degrees of freedom
Residual deviance:  773.89  on 1358  degrees of freedom
AIC: 797.89

Number of Fisher Scoring iterations: 7
```

The null deviance is the same for both the logistic regression using the original data and the logistic regression using the coordinates. This is obvious, since the null deviance shows how well the response variable is predicted by a model that includes only the intercept and excludes all other therms, and both models have the same response variables. Also the residual deviance and the AIC-value are very similar for both models. The big difference can be seen in the variables: Both models do not have the same significant variables. `Si_XRF` is highly significant in the model with coordinates, but does not describe the response variable significantly in the model with the original data. The original data are not represented in the usual Euclidean space, therefore the results and the interpretation of these results can be misleading.

For the last model, the compositional data are clr transformed by the `cenLR()` function implemented in R and available in the R-package robCompositions. The clr transformed compositional data set leads to the following result:

```
Call:
glm(y~.,data=data.frame(y=north,X=cenLR(X1)$ x.clr), family=binomial);

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9093  -0.4347  -0.1156   0.1615   2.9469

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  20.8193     3.0824   6.754 1.44e-11 ***
X.Al_XRF    -11.8233     1.1048 -10.702  < 2e-16 ***
X.Ca_XRF     -1.4988     0.3604  -4.158 3.20e-05 ***
X.Fe_XRF      1.8212     0.5475   3.326  0.00088 ***
X.K_XRF       2.6641     0.4524   5.889 3.89e-09 ***
```

```
X.Mg_XRF     -0.7371     0.3568  -2.066  0.03885 *
X.Mn_XRF     -2.5084     0.3565  -7.036 1.97e-12 ***
X.Na_XRF      4.3070     0.3510  12.271  < 2e-16 ***
X.P_XRF      -1.3767     0.4405  -3.125  0.00178 **
X.Si_XRF     -2.4000     0.3977  -6.035 1.59e-09 ***
X.Ti_XRF     -1.3957     0.4374  -3.191  0.00142 **
X.LOI             NA         NA      NA       NA
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1559.19  on 1369  degrees of freedom
Residual deviance:  775.85  on 1359  degrees of freedom
AIC: 797.85

Number of Fisher Scoring iterations: 7
```

Null deviance, residual deviance and AIC are the same for both models with ilr and clr transformed compositional data. The significance of the single variables is different though. The coefficients for the coordinate representing `LOI` is not defined by the model because of singularities. The clr (centered logratio) transformation is an isometric transformation and is defined by the logratio to the geometric mean of all variables (Filzmoser and Hron, 2009). The composition is mapped to a D-dimensional Euclidean vector subspace; thus, the transformation is not injective. The resulting transformed data are always singular.
A classical way to get rid of the singularity of the classical covariance matrix of compositional data is to erase one component. Certainly this procedure is not permutation-invariant, as results will largely depend on which component is erased (Pawlowsky-Glahn et al., 2007).

Summarizing we can state, that the isometric logratio (ilr) transformation has the most preferable properties among the transformations selected, as it avoids data singularity (see Hron et al. (2012)). With the application of the ilr transformation we can handle the fact, that compositional data follow the Aitchison geometry on the simplex and not the usual Euclidean geometry.

Now we select the final model using the ilr transformed variables by estimating the performance of different models in order to choose the best one. We create 100 training and test data sets and fit a model for every training data set with the help of the Akaike Information Criterion (AIC). The model selection is based on a stepwise selection procedure similar to backward selection in linear regression. This backward AIC elimination procedure starts with all predictors in the model and removes predictors until the AIC stops decreasing. The basic structure of this procedure is:

1. Remove each variable currently in the model, fit the model and calculate the AIC.
2. Choose the variable that leads to the model with the smallest AIC.
3. If the AIC of this model without this variable is lower than the AIC of the model containing the variable, drop the variable and go back to step 1. Otherwise, do not drop the variable and stop.

For our randomly selected 100 training data sets out of the GEMAS project with the `glmCoDaX()` we get in total 100 different models. In order to evaluate the performance of the single models, we use the corresponding test data sets to make a prediction for the response variable and compare the prediction with the real observation. In that way we can understand the accuracy of a prediction. The `predict.glm()` function in R can be used to predict the north - south localization of the observations. We supply the test data set to the `predict()` function, then probabilities are computed for the test data set. In order to make a prediction as to whether the observation is localized in northern or southern Europe, the predicted probabilities have to be converted into a logical operator with TRUE/FALSE elements. The `table()` function can be used to produce a confusion matrix in order to determine how many observations were correctly or incorrectly classified. The `mean()` function can be used to compute the fraction of observations for which the prediction was correct. In our case, logistic regression correctly predicted the localization of the test observation sites between 85,28% and 90,67% of the time. The test error rates of all models lie therefore between 9.33% and 14.72%. In Figure 5.3 the distribution of the rate of proper classified observations is shown. Figure 5.4 depicts the distribution of the test error rate for all 100 models. We see, that all models work quite well.

The 100 different fitted models can be grouped into the following 7 model classes:

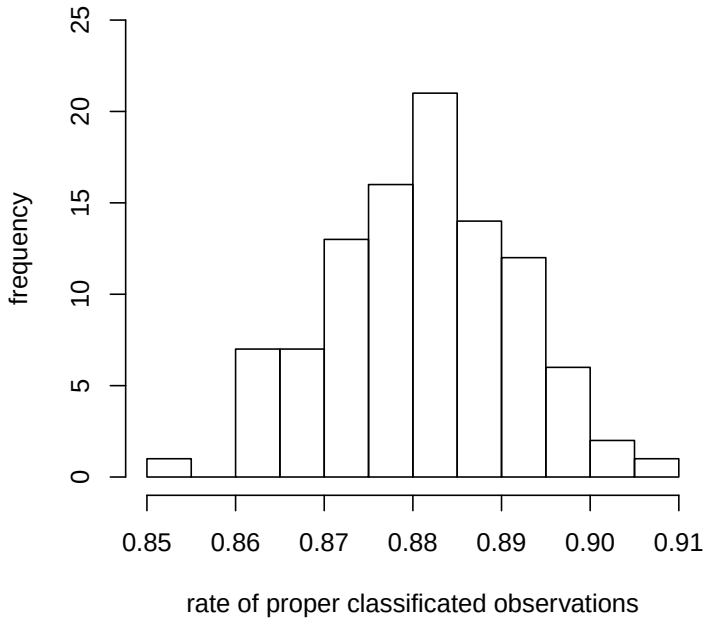

Figure 5.3: Histogramm of proper classificated observations.

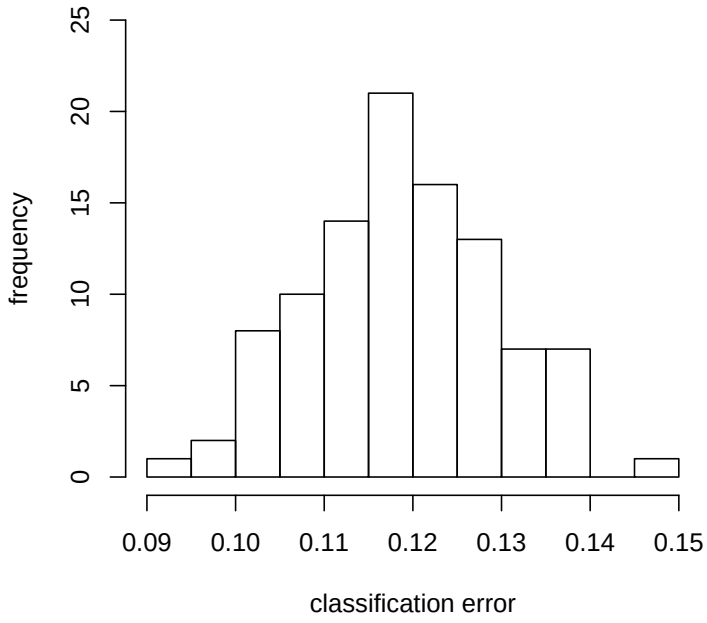

Figure 5.4: Histogramm of classification error for test data.

```
Model 1
NORTH ~ Al_XRF + Ca_XRF + Fe_XRF + K_XRF + Mg_XRF + Mn_XRF + Na_XRF +
P_XRF + Si_XRF + LOI

Model 2
NORTH ~ Al_XRF + Ca_XRF + Fe_XRF + K_XRF + Mg_XRF + Mn_XRF + Na_XRF +
Si_XRF + Ti_XRF + LOI

Model 3
NORTH ~ Al_XRF + Ca_XRF + Fe_XRF + K_XRF + Mg_XRF + Mn_XRF + Na_XRF +
Si_XRF + LOI

Model 4
NORTH ~ Al_XRF + Ca_XRF + Fe_XRF + K_XRF + Mn_XRF + Na_XRF + Si_XRF +
Ti_XRF + LOI

Model 5
NORTH ~ Al_XRF + Ca_XRF + Fe_XRF + K_XRF + Mn_XRF + Na_XRF + Si_XRF + LOI

Model 6
NORTH ~ Al_XRF + Fe_XRF + K_XRF + Mn_XRF + Na_XRF + Si_XRF + Ti_XRF + LOI

Model 7
NORTH ~ Al_XRF + Fe_XRF + K_XRF + Mn_XRF + Na_XRF + Si_XRF + LOI
```

The barplot in Figure 5.5 depicts the single model frequency. Model 3 with the ilr transformed variables `Al_XRF, Ca_XRF, Fe_XRF, K_XRF, Mg_XRF, Mn_XRF, Na_XRF, Si_XRF` and `LOI` is the most selected by the backward AIC elimination procedure. 45 training data sets resulted in model 3.
The mean test error rate for these 45 versions of model 3 amounts to 12,00%. On average these 45 models predicted correctly 88,00% of all observations.

We can also perform statistical tests to determine if subsets of the regression coefficients differ from zero. To perform this test we select the model out of the above 45 models with the minimal AIC.

`glmCoDaX`

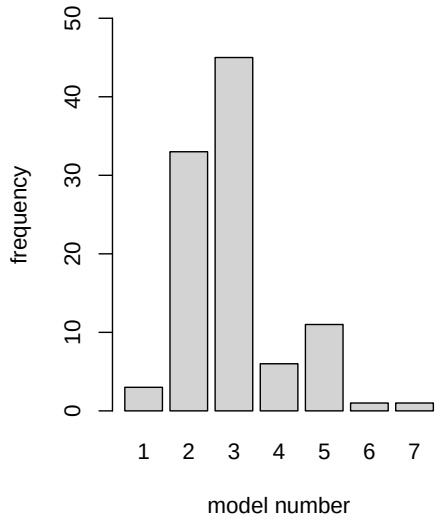

Figure 5.5: Barplot of model frequency

```
Call:
glm(formula = y ~ ., family = family, data = d)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.68021  -0.40777  -0.07923   0.21395   2.77996

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  18.5426     2.7634   6.710 1.94e-11 ***
X.Al_XRF    -11.9704     1.0123 -11.825  < 2e-16 ***
X.Ca_XRF     -0.5833     0.1855  -3.145 0.001660 **
X.Fe_XRF      3.0305     0.5109   5.932 2.99e-09 ***
X.K_XRF       4.2233     0.4561   9.260  < 2e-16 ***
X.Mg_XRF      0.9666     0.3328   2.904 0.003682 **
X.Mn_XRF     -1.7847     0.2944  -6.061 1.35e-09 ***
X.Na_XRF      6.1263     0.4036  15.178  < 2e-16 ***
```

```
X.Si_XRF     -0.9274     0.2444  -3.794 0.000148 ***
X.LOI         0.9192     0.2276   4.038 5.39e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1580.13  on 1369  degrees of freedom
Residual deviance:  746.13  on 1361  degrees of freedom
AIC: 764.13

Number of Fisher Scoring iterations: 8
```

For this model, the ilr transformed variables `Al_XRF, Fe_XRF, K_XRF, Mn_XRF, Na_XRF`, `Si_XRF` and `LOI` are highly significant.
For this special model, logistic regression correctly predicted the localization of the test observation 86,44% of the time. The test error rate is therefore 13,56%.

We compare this model with the full model by a likelihood ratio test. The full model for the same training data set is the following:

```
glmCoDaX

Call:
glm(formula = y ~ ., family = family, data = d)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.69449  -0.41085  -0.07658   0.20428   2.77418

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  19.3710     3.1774   6.096 1.08e-09 ***
X.Al_XRF    -12.0758     1.0659 -11.329  < 2e-16 ***
X.Ca_XRF     -0.5913     0.1856  -3.186 0.001443 **
X.Fe_XRF      3.0314     0.5429   5.583 2.36e-08 ***
X.K_XRF       4.2958     0.4787   8.974  < 2e-16 ***
```

```
X.Mg_XRF      0.9404     0.3300   2.849 0.004382 **
X.Mn_XRF     -1.7496     0.2936  -5.958 2.55e-09 ***
X.Na_XRF      6.1632     0.4250  14.503  < 2e-16 ***
X.P_XRF      -0.2365     0.2675  -0.884 0.376587
X.Si_XRF     -0.9248     0.2793  -3.311 0.000930 ***
X.Ti_XRF      0.1029     0.3970   0.259 0.795477
X.LOI         1.0444     0.2722   3.838 0.000124 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1580.13  on 1369  degrees of freedom
Residual deviance:  745.31  on 1359  degrees of freedom
AIC: 767.31

Number of Fisher Scoring iterations: 8
```

Our reduced and the full model are nested models. Both models can be compared using the deviance D. Considering the full model

$$logit(p(x_1, x_2, ..., x_p)) = \beta_0 + \beta_1 x_1 + ... + \beta_q x_q + \beta_{q+1} x_{q+1} + ... + \beta_p x_p,$$

we want to test the null hypothesis

$$H_0 : \ \beta_{q+1} = ... = \beta_p = 0,$$

versus the alternative hypothesis that at least one of these coefficients differs from zero. If H_0 is true, then the regressor variables $x_{q+1}, ..., x_p$ are redundant in the full model and can be dropped. In order to test H_0 in practice, the respective deviances for both models were compared. The test statistic is:

$$\mathbf{X}^2 = D_{reduced} - D_{full}.$$

Both deviances in the above equation involve the evaluation of the log-likelihood for the saturated model, as the deviance D is defined as

$$D = 2[\text{log-likelihood(saturated model)} \ - \ \text{log-likelihood(proposed model)}].$$

The saturated model has as many parameters as observations. When we take in the test statistic the differences of the deviances (reduced - full), the log-likelihood for the saturated model cancels out. Therefore the test statistic $\mathbf{X^2}$ can be written as

$$\mathbf{X}^2 = 2[\text{log-likelihood(full model)} - \text{log-likelihood(reduced model)}] :$$

This test statistic is sometimes referred to as the log-likelihood ratio statistics.
If H_0 is true, then the test statistic $\mathbf{X^2}$ has an approximate chi-squared distribution whose degrees of freedom is equal to the difference in the number of parameters between the full and reduced models: $p - q$. When testing at a level of significance α, then we reject H_0 if $\mathbf{X^2} > \chi_{\alpha,p-q}$, the α critical value of the chi-squared distribution on $p - q$ degrees of freedom.
The residual deviance for the full model is 745.31 on 1359 degrees of freedom, the residual deviance for the reduced model 746.13 on 1361 degrees of freedom. The value of the log-likelihood ratio statistics is $\mathbf{X^2} = 0.82$. Since the full and reduced models differ by 2 parameters, we can compare this test statistic to a chi-squared distribution on 2 degrees of freedom. The p-value for this test is $p = 0.6637$. Thus we conclude, that there is insufficient evidence that the coefficients regarding the variables `P_XRF` and `Ti_XRF` differ from zero. We can select the model without these two coefficients.

Finally we are interested in the information about which observations are how often misclassified by the 45 versions of model 3, which we selected as final model. The misclassification is tested for every model with the full data sets. In Figures 5.6 and 5.7 the results are depicted. It is interesting to note, that almost all observations in the area of the Baltic states were misclassified.

5.2 Count data, Poisson regression and log-linear model

In this section we apply generalized linear models with a Poisson random variable and a compositional explanatory data set.

5.2.1 Biomonitoring campaign in Czech Republic

The project "Biogeochemical exploration of forests as a base for the long-term landscape exploitation in the Czech Republic" (CZ0074) was implemented between July 2008 and March 2011 by the Silva Tarouca Research Institute for Landscape and Ornamental

Figure 5.6: Misclassification of observations. Black color means misclassification 100% of the time; white color no misclassification.

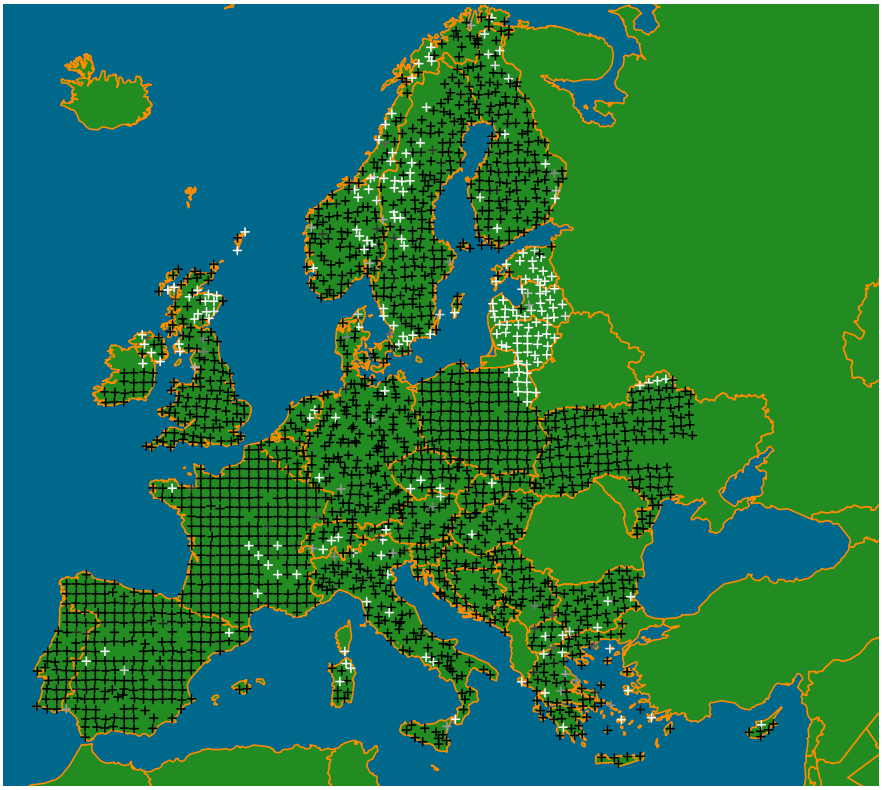

Figure 5.7: Misclassification of observations. White color means misclassification 100% of the time; black color no misclassification.

Gardening, Public Research Institution (abbrev VÚKOZ) in co-operation with the Geological Survey of Norway (Norges geologiske undersøkelse, abbrev NGU). The project was supported by a grant from Norway through the Norwegian Financial Mechanism and the Revolving Fund of the Ministry of the Environment of the Czech Republic.
The project focuses were the acquirement of complex information about forests biogeochemistry and especially anomalies in the concentration distributions of toxic and high-risk elements in chosen bio-indicators of environmental quality in the Czech Republic.
Samples of big red stem moss (Pleurozium schreberi), wavy hairgrass (Deschampsia flexuosa) and annual and biennial Norway spruce needles (Picea abies) receiving chemical elements mainly from atmospheric deposition, forest floor humus and upper and lower mineral forest soils (depleted and enriched B horizons) were collected in parallel over the territory of the Czech Republic at an average sample density of about 1 site/300 km^2. In total 250-280 plots evenly spread over the Czech territory were selected (see Figure 5.8). The sample sites are part of the "permanent monitoring plots" used for the European and national moss monitoring campaigns (UNECE ICP-Vegetation).

Concentrations of in total 39 elements were determined using inductively-coupled atomic-emission spectrometry (ICP-AES) and dynamic reaction cell inductively-coupled plasma mass spectrometry (DCR ICP-MS, Perkin Elmter) instruments. These elements were Ag, Al, As, Ba, Be, Bi, (Ca), Cd, Ce, Co, Cr, Cs, Cu, Fe, Ga, Ge, Hg, (K), La, Li, (Mg), Mn, Mo, (Na), Nd, Ni, Pb, Pr, Rb, S, Sb, Se, Sn, Sr, Th, Tl, U, V, Y and Zn (the major nutrients Ca, K, Mg and Na were not analyzed in moss). The quality of the analytical results was checked through analyzes of standard reference plant materials.
The aim of the project was to synthesize obtained data of relatively high concentrations of toxic and high-risk elements and deficiencies of biologically active elements in investigated bio-indicators. Potential risks for forests ecosystems and human health in case of determination of extreme accumulation of the monitored elements should be evaluated. The project should supply the information to relevant public authorities as a basis for possible corrective actions. This would improve the quality of decision-making processes in the sphere of long-term land use, nature conservation and health protection in the Czech Republic.

5.2.2 Data set

The data set of the project "Biogeochemical exploration of forests as a base for the long-term landscape exploitation in the Czech Republic" (CZ0074) consists in the con-

Figure 5.8: Location of study sites within Czech Republic.

centration of several chemical elements for every bioindicator selected. The sample number amounts to 254 for Norway spruce needles, 249 for grass, 259 for humus and 280 for moss. For every plot there has been collected also data like elevation, annual precipitation or geomorphology. Additionally data about nearest distance to road and the respective average number of cars and trucks (in 24 hours) traveling on certain roads in the year 2005 is provided in the project. The distance to the nearest road is indicated in metres. Minimum distance is null, maximum distance 1.959 m. No information about number of cars and trucks is indicated if there is no road up to 2000 m from the sampling plot or no data about traffic at the given road segments - usually very small roads with a negligible traffic intensity. In total 116 plots dispose of an information about traffic volume.

5.2.3 Traffic induced pollution in Czech Republic

The project "Biogeochemical exploration of forests as a base for the long-term landscape exploitation in the Czech Republic" (CZ0074) showed, that it can be possible to find an appropriate bioindicator for a certain task (atmospheric dust level, urban contamination,...) and a certain group of elements. Results of the study let suppose, that an universal bioindicator reflecting anthropogenic input 1:1 does not exist. One of the best bioindicators for monitoring atmospheric contamination at a regional-scale appears to be moss. Moss can not provide a reflection of the complete atmospheric deposition but rather a picture that is biased towards certain elements like V, Pb and Sb. Thus moss may preferably reflect traffic emissions (Suchara et al., 2011).
The aim of our investigation is the identification of relations between traffic volume and chemical composition of moss. The traffic volume is included in the investigation via the count data representing the average number of cars and trucks traveling on certain roads. The chemical composition of moss is available as compositional data set including the following chemical elements: Ag, Al, As, Ba, Be, Bi, Cd, Ce, Co, Cr, Cs, Cu, Fe, Ga, Hg, La, Li, Mn, Mo, Nd, Ni, Pb, Pr, Rb, S, Sb, Se, Sn, Sr, Th, Tl, U, V, Y, Zn. Using both data sets, we try to fit a generalized linear model that describes a relation between them. We use the whole data set and select the final model via backward AIC elimination procedure. The R-function `glmCoDaX()` is used. The input for a Poisson GLM with log link and our data is given by:

```
glmCoDaX(X,y,poisson),
```

with X a matrix representing 280 compositions of ilr transformed concentrations of 35 chemical elements and y a vector with count data representing traffic volume. As in

the previous section about binomial regression already indicated, the ilr transformed variables are ratios and therefore dimensionless numbers, which have not an obvious meaning.
After application AIC elimination procedure only the variable *Al* is eliminated. All the other variables are highly significant. The function delivers the following result:

```
glmCoDaX

Call:
glm(formula = y ~ ., family = family, data = d)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-107.709   -33.769    -9.557    18.291   117.220

Coefficients:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept) 18.168606   0.283059   64.187  < 2e-16 ***
X.Ag        -1.005612   0.012062  -83.372  < 2e-16 ***
X.As        -0.569064   0.011047  -51.513  < 2e-16 ***
X.Ba         0.283509   0.008681   32.657  < 2e-16 ***
X.Be         0.220710   0.009904   22.285  < 2e-16 ***
X.Bi        -0.371845   0.011668  -31.869  < 2e-16 ***
X.Cd        -0.076522   0.009672   -7.912 2.53e-15 ***
X.Ce        -2.207928   0.091325  -24.177  < 2e-16 ***
X.Co         0.256049   0.010997   23.283  < 2e-16 ***
X.Cr         0.054836   0.015149    3.620 0.000295 ***
X.Cs        -0.204661   0.005918  -34.581  < 2e-16 ***
X.Cu         2.573601   0.020032  128.476  < 2e-16 ***
X.Fe         1.638624   0.015247  107.471  < 2e-16 ***
X.Ga         0.288644   0.028879    9.995  < 2e-16 ***
X.Hg         0.588192   0.012569   46.798  < 2e-16 ***
X.La         5.914467   0.070625   83.745  < 2e-16 ***
X.Li        -0.644279   0.017682  -36.438  < 2e-16 ***
X.Mn        -0.080779   0.004695  -17.206  < 2e-16 ***
X.Mo        -0.871924   0.014192  -61.438  < 2e-16 ***
X.Nd        -8.936795   0.138837  -64.369  < 2e-16 ***
```

```
X.Ni        -0.517476   0.014043  -36.850  < 2e-16 ***
X.Pb         0.304473   0.012587   24.190  < 2e-16 ***
X.Pr         3.835915   0.166693   23.012  < 2e-16 ***
X.Rb        -0.100429   0.005575  -18.015  < 2e-16 ***
X.S         -2.817398   0.022305 -126.312  < 2e-16 ***
X.Sb         0.945839   0.015101   62.633  < 2e-16 ***
X.Se         0.303421   0.011450   26.499  < 2e-16 ***
X.Sn        -0.300137   0.021134  -14.201  < 2e-16 ***
X.Sr        -0.097788   0.010273   -9.519  < 2e-16 ***
X.Th         1.330992   0.023146   57.503  < 2e-16 ***
X.Tl         0.376190   0.006262   60.072  < 2e-16 ***
X.U         -0.127732   0.012328  -10.361  < 2e-16 ***
X.V         -1.621152   0.025010  -64.819  < 2e-16 ***
X.Y          0.913852   0.027569   33.147  < 2e-16 ***
X.Zn         0.722205   0.013097   55.143  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 314314  on 115  degrees of freedom
Residual deviance: 183118  on  82  degrees of freedom
AIC: 184250

Number of Fisher Scoring iterations: 6
```

In order to evaluate the performance of the model, we use the same compositional data set to make a prediction for the response variable and compare the prediction with the real observation. A comparison between the really observed data and the predicted data provides the following result, where the variable `y` stands for the observations, and the variable `pred` for the predictions:

```
> summary(y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   86.0   730.5  1557.0  2681.0  3438.0 21230.0

> summary(pred)
```

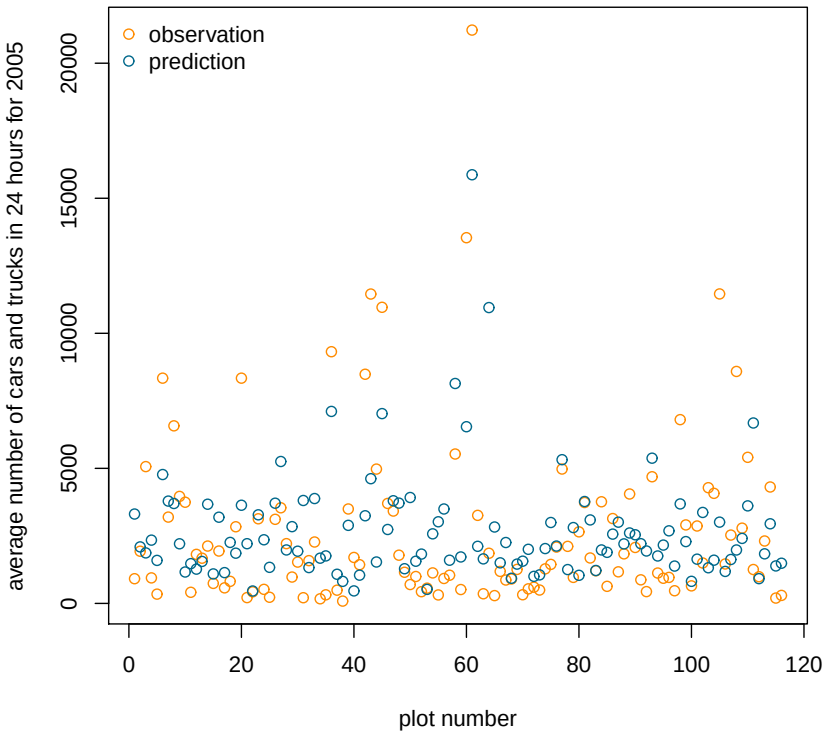

Figure 5.9: Plot of observed and predicted data regarding traffic volume for 116 sample plots over Czech Republic

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  459.5  1526.0  2142.0  2681.0  3254.0 15870.0
```

Figure 5.9 depicts the results for the single plots (number of plots $n = 116$).
The mean is the same for the predicted and the observed values. The model obviously is not ideal: the data are clearly overdispersed as it underestimates the variance in the data. Dispersion is an important concept in the analysis of discrete data. The dispersion parameter describes if variance is greater than the mean (overdispersed) or lower (underdispersed). McCullagh and Nelder (1989) say, that overdispersion is the rule rather than the exception; underdispersion is rare in practice. In GLM for every distribution a certain ratio between mean and variance is assumed. For Poisson distribution the variance is identical to the mean. If the Poisson model is a good fit to the data, then it follows that the deviance should be roughly equal to the deviance degrees of freedom which is

the sample size n minus the number of estimated coefficients. If the deviance greatly exceeds the degrees of freedom, then that is an indication of an overdispersion problem. In R, the `glm()` function uses the residual deviance. The problem of overdispersion can be caused by the specification of the model. Another reason overdispersion can occur is that a different count distribution is needed to model the data. Probably in our case a negative Binomial or quasi-Poisson model would be more appropriate. Further data investigation is in any case needed.

6 Conclusion

In recent years linear regression with compositional data has been investigated (Hron et al., 2010, 2012). Hron et al. (2012) used an approach based on the isometric logratio (ilr) transformation. The resulting model turned out to be easy to handle with parameter estimation like in usual linear regression. In addition they used the ilr variables for inference statistics in order to obtain appropriate interpretation of the model.
Generalized linear models with compositional data however is a rather missing topic in statistics so far. The aim of this study thus was to implement generalized linear models with compositional data sets.
Generalized linear models are a generalization of the classical linear regression. The term "generalized linear models" - introduced by Nelder and Wedderburn (1972) and popularized by McCullagh and Nelder (1989) - refers to a larger class of models. In these models, the response variable y_i is assumed to follow an exponential family distribution with mean μ_i, which is assumed to be some linear function of $\mathbf{x}_i^T\boldsymbol{\beta}$.
In these models compositional data can not be used directly. Compositional data consist of a vector, where the single components are positive, describe the parts of some whole and carry only relative information. These data are not represented in the usual Euclidean space, but follow the so-called Aitchison geometry on the simplex. The interpretation of the model, and also the inference statistics for the estimated parameters are only valid if the data are represented in the appropriate Euclidean space.
We apply a transformation on the compositional explanatory data set in order to express the original compositions in orthonormal coordinates with respect to the Aitchison geometry. The construction of such orthonormal coordinates is crucial for the interpretability of the results. The isometric logratio (ilr) transformation according to Hron et al. (2012) is applied. Here the first ilr coordinate explains the compositional part of interest. In application of this special ilr transformation a meaningful interpretation of the unknown parameters and the inference statistics is possible. Anyway, isometric logratio transformation shows the most preferable properties among the transformations selected in this study. The log transformed original data are not represented in the usual Euclidean space, therefore the results and the interpretation of these results can be misleading. The centered logratio (clr) transformation is not appropriate because it results

in singularity.
We implemented generalized linear models with compositional data sets in R. The respective R-Code for this new function called `glmCoDaX()` can be found in the appendix of this work.
Generalized linear models are most commonly used to model binary or count data. Therefore we focused on models for these types of data. We used a compositional data set resulting from the GEMAS (Geochemical mapping of agricultural and grazing land soils) project, a cooperation project between EuroGeoSurveys and Eurometaux, to implement a model for binary data. The compositional data set consisted in the total concentration of 10 major elements of a soil sample plus Loss on Ignition for 2056 observations across 33 European countries. The response variable consisted in a binary variable expressing whether the observation is localized in northern or southern Europe. Our model, a binomial GLM with logit link function correctly predicted the localization of the test observation sites between 85,28 % and 90,67 % of the time. Only evident discrepancy was found in the area of the Baltic state. Almost all observations in this area were misclassified by the model. Apart from this discrepancy a strong relation between localization of the observation sites within northern or southern Europe and the chemical soil composition was found. A similar result was found in the GEMAS project, where large difference in the concentration of many chemical elements between the young soils from northern Europe and the older and more weathered southern European soil could be observed.
Moreover we used a compositional data set from the project "Biogeochemical exploration of forests as a base for the long-term landscape exploitation in the Czech Republic" to implement a model for a Poisson random variable and a log-linear model. The aim of our investigation was the identification of relations between traffic volume and chemical composition of moss. The resulting model was not ideal, the data is clearly overdispersed. The problem of overdispersion can be caused by the specification of the model regarding for example omitted variables and functional forms. Another reason overdispersion can occur is that a different count distribution is needed to model the data. Probably a negative Binomial or quasi-Poisson model would lead to results that are more accurate.

7 Appendix: R-Code

```
glmCoDaX <- function(X,y,family){

# delivers appropriate inference for generalized linear models of y
# on a  compositional matrix X

# At first a classical generalized linear model is implemented
# to build the structure of the output, which is then filled with
# the results of isomLR generalized linear models.
d <- data.frame(y=y,X=X)
suppressWarnings(glmcla <- glm(y~.,data=d,family=family))
glmcla.sum <- summary(glmcla)

# isomLR generalized linear models
require(robCompositions)
ilrglm.sum <- glmcla.sum
for (j in 1:ncol(X)){
  Zj <- -robCompositions::isomLR(cbind(X[,j],X[,-j]))
  dj <- data.frame(y=y,Z=Zj)
  res <- glm(y~.,data= dj,family=family)
  res.sum <- summary(res)
  if (j==1){
    ilrglm.sum$coefficients[1:2,] <- res.sum$coefficients[1:2,]
    ilrglm.sum$deviance.resid<- res.sum$deviance.resid
    ilrglm.sum$deviance <- res.sum$deviance
    ilrglm.sum$null.deviance <- res.sum$null.deviance
    ilrglm.sum$aic <- res.sum$aic
ilrglm.sum$df.residual <- res.sum$df.residual
  }
  else{
    ilrglm.sum$coefficients[j+1,] <- res.sum$coefficients[2,]
```

```
  }
}
list(glmCoDaX=ilrglm.sum)
}
```

List of Figures

List of Tables

Bibliography

Agresti, A., 2002. Categorical Data Analysis. Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., New York, NY [u.a.].

Aitchison, J., 1986. The statistical analysis of compositional data. (2010 print of 2003 reprint by the blackburn press) ed., Chapman and Hall, London - New York.

Dobson, A.J., 2002. An introduction to generalized linear models. Texts in statistical science. 2. ed. ed., Chapman & Hall, CRC, Boca Raton, Fla. [u.a.].

Egozcue, J., Pawlowsky-Glahn, V., 2005. Groups of parts and their balances in compositional data analysis. Mathematical Geology 37, 795–828.

Egozcue, J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. Mathematical Geology 35, 279–300.

Egozcue, J.J., Pawlowsky-Glahn, V., 2011. Basic concepts and procedures, in: Pawlowsky-Glahn, V., Buccianti, A. (Eds.), Compositional Data Analysis: Theory and Applications. John Wiley & Sons, Ltd, pp. 12–28.

EuroGeoSurveys, 2015. The Geological Surveys of Europe. URL: `www.eurogeosurveys.org`.

Filzmoser, P., Hron, K., 2009. Correlation analysis for compositional data. Mathematical Geosciences 41, 905–919.

Filzmoser, P., Hron, K., Reimann, C., 2009. Univariate statistical analysis of environmental (compositional) data: Problems and possibilities. Science of the Total Environment 407, 6100–6108.

Filzmoser, P., Hron, K., Reimann, C., 2010. The bivariate statistical analysis of environmental (compositional) data. Science of The Total Environment 408, 4230 – 4238.

Fox, J., 2008. Applied regression analysis and generalized linear models. 2. ed. ed., Sage Publ., Los Angeles, Calif. [u.a.].

GEMAS, 2015. Project Homepage. URL: `http://gemas.geolba.ac.at`.

Hron, K., Filzmoser, P., Thompson, K., 2012. Linear regression with compositional explanatory variables. Journal of Applied Statistics 39, 1115 – 1128.

Hron, K., Templ, M., Filzmoser, P., 2010. Imputation of missing values for compositional data using classical and robust methods. Computational Statistics & Data Analysis 54, 3095 – 3107.

Martín-Fernández, J.A., Palarea-Albaladejo, J., Olea, R.A., 2011. Dealing with zeros, in: Pawlowsky-Glahn, V., Buccianti, A. (Eds.), Compositional Data Analysis: Theory and Applications. John Wiley & Sons, Ltd, pp. 43–58.

Mateu-Figueras, G., Pawlowsky-Glahn, V., Egozcue, J.J., 2011. The principle of working on coordinates, in: Pawlowsky-Glahn, V., Buccianti, A. (Eds.), Compositional Data Analysis: Theory and Applications. John Wiley & Sons, Ltd, pp. 29–42.

McCullagh, P., Nelder, J.A., 1989. Generalized linear models. Monographs on statistics and applied probability ; 37. 2. ed. ed., Chapman and Hall, London [u.a.].

Nelder, J.A., Wedderburn, R.W.M., 1972. Generalized linear models. Journal of the Royal Statistical Society. Series A (General) 135, 370–384.

Pawlowsky-Glahn, V., Buccianti, A.E., 2011. Compositional Data Analysis: Theory and Applications. John Wiley & Sons, Ltd.

Pawlowsky-Glahn, V., Egozcue, J., 2002. Blu estimators and compositional data. Mathematical Geology 34, 259–274.

Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R., 2007. Lecture Notes on Compositional Data Analysis. URL: `http://dugi-doc.udg.edu//handle/10256/297`.

Reimann, C., Filzmoser, P., Fabian, K., Hron, K., Birke, M., Demetriades, A., Dinelli, E., Ladenberger, A., 2012. The concept of compositional data analysis in practice - total major element concentrations in agricultural and grazing land soils of europe. Science of The Total Environment 426, 196 – 210.

Suchara, I., Sucharova, J., Hola, M., Reimann, C., Boyd, R., Filzmoser, P., Englmaier, P., 2011. The performance of moss, grass, and 1- and 2-year old spruce needles as bioindicators of contamination: A comparative study at the scale of the czech republic. Science of the Total Environment 409, 2281–2297.

Templ, M., Hron, K., Filzmoser, P., 2015. robCompositions: Robust Estimation for Compositional Data. Manual and package, version 1.9.1. URL: `http://cran.r-project.org/web/packages/robCompositions/index.html`.