course AB_450071

Statistics and Data Analysis (Part-II: Data Analysis)

Faculty of Science

Vrije Universteit Amsterdam

for Aardwetenschappen BSc

for Aarde, Economie en Duurzaamheid BSc

Versie 2024

Niels J. de Winter

Modified after a previous version by J. van Huissteden

Table of Contents

1. VOORWOORD IN HET NEDERLANDS

Leerdoelen van dit cursusonderdeel

Welkom bij het tweede deel van de cursus **Statistiek en Data Analyse**. In dit deel van de cursus ligt de focus op het beschrijven van trends en patronen in datasets op een statistisch verantwoorde manier. We gebruiken hierbij de technieken die we in deel één van de cursus (Statistiek) hebben geleerd op een grotere schaal om datasets te leren begrijpen. Tijdens deze cursus behandelen we een aantal verschillende soorten "tools" voor data analyse. Het doel van deze cursus is om deze methodes te leren kennen, toepassen en de resultaten te leren interpreteren. We hebben daarom gekozen om in deze cursus gebruik te maken van korte colleges waarin de theoretische basis van de data analyse tools wordt uitgelegd met enkele voorbeelden. Daarnaast bieden we computerpractica aan waarin de kans wordt geboden om de nieuw geleerde technieken toe te passen op concrete datasets, meestal met een oorsprong in de Aardwetenschappen. Deze syllabus is daarnaast bedoeld als naslagwerk om de theorie nog eens door te kunnen nemen.

Belang van oefening

Omdat de focus van de cursus ligt op het leren begrijpen en toepassen van data analyse methoden zijn de computerpractica van essentieel belang voor je begrip van de stof. De ervaring leert dat actieve deelname aan de computerpractica de kans van slagen voor deze cursus significant verhoogt. De examens voor het data analyse deel van deze cursus omvatten deels vragen over de theorie en deels toepassingsvragen die erg lijken op de opdrachten die tijdens het computerpracticum behandeld worden. Met alleen het bestuderen van de theorie is het daardoor erg lastig om ook deze vragen goed te beantwoorden.

Computerpractica en het gebruik van Python en Jupyter

De computerpractica worden aangeboden in de Notebook-omgeving **Jupyter** waarbinnen de programmeertaal **Python** wordt gebruikt voor de berekeningen. Voor sommige studenten is dit een eerste aanraking met Jupyter of Python. Vandaar dat dit deel van de cursus wordt ingeleid met een college en een werkgroep waarin het gebruik van Jupyter en Python centraal staan. Deze syllabus behandelt geen technische aspecten van Python of Jupyter, maar tijdens de cursus worden online naslagwerken aangeboden waarin meer informatie over het gebruik van Python en Jupyter staat. Daarnaast is er een levendige online gemeenschap van gebruikers van Python en Jupyter, en worden studenten aangemoedigd om online naar oplossingen voor problemen te zoeken buiten de contacturen.

Taal

De voertaal van deze cursus is Nederlands. Echter hebben we ervoor gekozen om de theorie in deze syllabus, met uitzondering van deze inleiding, en een deel van het overige cursus materiaal (slides, opdrachten, etc.) in het Engels aan te bieden. De reden hiervoor is dat de voertaal van de hedendaagse wetenschap Engels is. Aangezien het leerdoel van deze cursus is om data analyse tools in een (Aard)wetenschappelijke context te kunnen toepassen, is het essentieel om de begrippen die we in deze cursus behandelen ook in het Engels te kennen. Werken in het Engels heeft als bijkomend voordeel dat de materialen ontewikkeld binnen deze cursus gemakkelijk (internationaal) overdraagbaar zijn (Open Educational Resources), en dat studenten eenvoudiger online hulp kunnen vinden door te zoeken in het Engels.

I. INTRODUCTION

These lecture notes introduce basic concepts in regression analysis, multivariate statistics, time series analysis and spatial analysis. We will start simple, with **bivariate analysis**, in which we analyze datasets that have two variables. For example, we may be interested in the relationship between atmospheric CO_2 concentrations and mean temperatures on earth.

We will then build on this basis to introduce **multivariate analysis**, which allows us to analyze datasets with more than two variables. Examples of such datasets within the geosciences include:

- Geochemical analysis of rocks where the concentrations of multiple elements has been measured in multiple samples.
- Research projects where a process depends om more than one variable, such as the decomposition of soil organic matter, which depends on soil temperature, soil moisture and soil acidity (pH), or real estate prices depending on various spatial economical variables.
- Datasets in which we want to classify samples based on their properties in more than two variables, such as phylogeny in which we want to relate organisms by reconstructing the tree of life.

Time series analysis is a special case of bivariate analysis in which the data we use consists of observations made on successive points in time. Examples of such datasets include:

- A series of meteorological observations such as air temperature or precipitation done over a certain period.
- The incidence of earthquakes of a certain magnitude in the Netherlands.
- A stratigraphic record in which the properties of successive dated rock or sediment layers are analyzed.

Spatial analysis can be seen as a special case of multivariate analysis, in which at least two of the variables define the place of a measurement and one or more other variables contain information of the measurement done in that place. Examples of such datasets could be:

- A geological map in which information about the subsurface structures are gathered across a certain area.
- A dataset containing information about the amount of rainfall measured at various weather stations.
- A dataset containing information about the median income of inhabitants of a country ordered by the municipality they live in.

Multivariate statistics and time series analysis in the earth sciences have some special problems which differ from similar procedures in other scientific disciplines. For instance, a geologic time series may be strongly different form a time series of stock prices, since the time axis is often less exactly known, or the time interval is variable. Earth Science data also often incomplete due to problems preventing us from collecting information from a certain time or place. These differences will be treated in the course. Furthermore, you often will find percentage data, which have special problems associated with them.

An important part of this course consists of computer practice. Data used in multivariate statistics, time series or spatial analysis usually of very large numbers of observations (“big data”). Thus, the

amount of calculation required to apply statistical analysis on these datasets is impractical without a computer with statistical software. The amount of data collected in modern geoscientific research is often large. Think of a satellite images, which consists of millions of image elements or pixels, each pixel containing information in several different wavelength bands of the electromagnetic spectrum.

Analysis of these huge amounts of data boils down to a few questions, all having the goal of finding some **order** and **patterns** in the data and to find out if the data tells us something about the processes that are behind the observations:

- Can I distinghuish a 'signal' in the data and seperate it from 'background noise'?
- Can I establish trends in the data and describe these with a mathematical formulation, e.g. variable *x* is related to variable *y* according to function *f*?
- Are some observations similar to others and can I discern groups of similar observations?
- Are variables related to each other, are there variables that vary in the same way?
- What can be the processes that cause the variation of the measured variables?
- Are observations made closer together in either time or space more similar than observations farther apart?

At the end of this course you will have a basic toolbox to answer such questions, and you have developed essential computer skills to perform data analysis. If you keep these questions in mind, this course will be more than just a series of tricks and manipulations with numbers. It will become a useful toolbox when you participate in your first research project or fieldwork and need to draw meaningful conclusions from a large bunch of numbers gathered in the field or the lab.

Throughout this syllabus, reference is made to the book by John C. Davis ("Statistics and Data Analysis in Geology"[1]). The lecture notes are meant to highlight the basic knowledge of data analysis that you will need as an Earth or environmental scientist. It is not a replacement of Davis' book, but gives extra explanation where necessary, to help you grasp the ideas behind the sometimes difficult to understand mathematical manipulations. For a more in-depth treatment of certain subjects, please refer to the book, but note that you do not need to have read the book to pass this course. Also, in these lecture notes some subjects are added that are not incorporated in the book by Davis.

[1] John C. Davis and Robert J. Sampson, *Statistics and Data Analysis in Geology*, vol. 646 (Wiley New York, 1986), https://www.kgs.ku.edu/Mathgeo/Books/Stat/ClarifyEq4-81.pdf.

II. CORRELATION

II.1 What is correlation?

Correlation is a method to test whether two variables **co-vary** (or "correlate") with each other. A positive correlation simply means that two variables vary in the same direction, while a negative correlation means that variables vary in opposite directions. A few examples of variables which are correlated are:

1. The number of ice cream cones sold increases when the average daily temperature increases (positive correlation).
2. The number of bee species in an area decreases when insecticide use in that area increases (negative correlation).
3. The number of people drowning in the sea increases with increasing outdoor temperatures (positive correlation).
4. The per-capita consumption of cheddar cheese in the USA increased at almost the same rate as the amount of energy generated by solar power in Haiti (positive correlation; see Figure 1).

II.2 Different reasons for correlation

As you can see from the examples above, a correlation between two variables does not necessarily entail a **causal relationship**. While in the first two examples you might be convinced that there is a causal relationship between the two variables, this seems hard to believe for examples 3 and 4. To prove that we have found a causal relationship, we need to do more than calculate the statistical correlation between two variables. Only by explaining how one variable influences the other can we demonstrate that the relationship is causal. For example 1, the explanation is that people have a stronger interest in eating ice cream to cool down when the weather is warm. In example 2, the insecticide used to deter pests from agricultural fields accidentally also kills wild bees (which are not generally considered pests).

In example 3, we are likely dealing with a **confounding variable**. A confounding variable is a variable which we did not measure, but which influences the two (or more) variables we measured in such a way that they correlate. In example 3, the confounding variable is likely to be the number of people going to the beach for a swim, which is higher when the weather is warmer, and which increases the chance that people drown. We might mistakenly interpret the correlation we found to mean that people drown more readily in warmer water, but this is not the case. When interpreting correlations, always watch out for confounding variables!

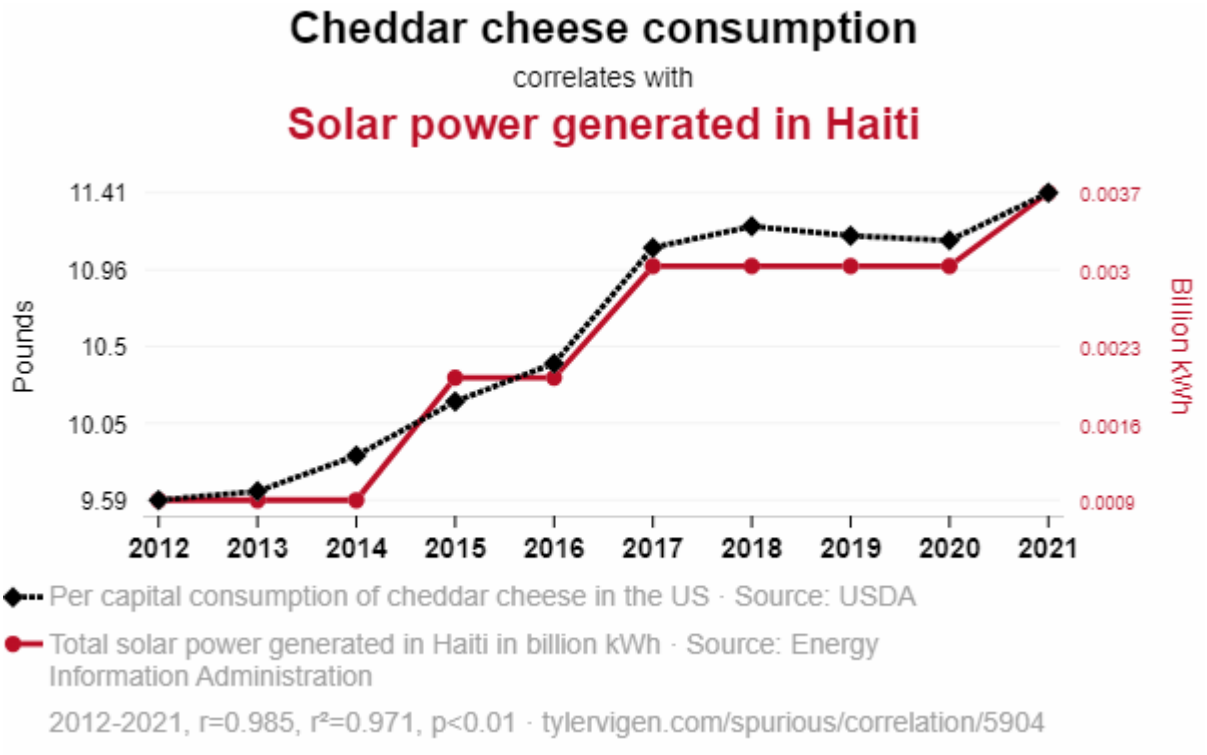

Figure 1: Time series of cheddar cheese consumption and solar power generation in Haiti show a strong correlation (© Tyler Vigen)

Example 4 (Figure 1) makes no logical sense at all, and the correlation is likely a result of **random chance**. It may seem tempting to look for a causal explanation between these variables because they co-vary so strongly, but you must keep in mind that if we compare many unrelated variables with each other, at some point we will discover accidental correlations that have no logical meaning. Tyler Vigen, author of the "spurious correlation" page[1] has raised this search for nonsensical correlations to an art form. Check out his website if you want to have a good laugh. By the way, please e-mail me if you think you can find a causal relationship between the variables in Figure 1, or even a confounding variable that indirectly links them. I am willing to buy you a good bottle of wine if you can convince me that this correlation demonstrates a causal effect!

II.3 Testing and quantifying correlations

If you want to know if two variables are correlated, the best first step is to create a scatterplot of the two variables (see Figure 2). However, as you will see, it can be hard to "eyeball" a correlation between two variables. To help us, we can use a *statistical coefficient* to test our correlation. The most commonly used correlation coefficient is **Pearson's correlation coefficient**, usually indicated with the letter "r". **Pearson's r** is the ratio between the *covariance* between two variables and the product of the standard deviations of the two variables (see Equation (1)).

$$Pearson's\ r = \frac{cov(X,Y)}{\sigma_X * \sigma_Y} = \frac{\sum(x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 * \sum(y_i - \bar{y})^2}} \quad (1)$$

[1] "Tyler Vigen's Personal Website," accessed March 22, 2024, https://tylervigen.com/.

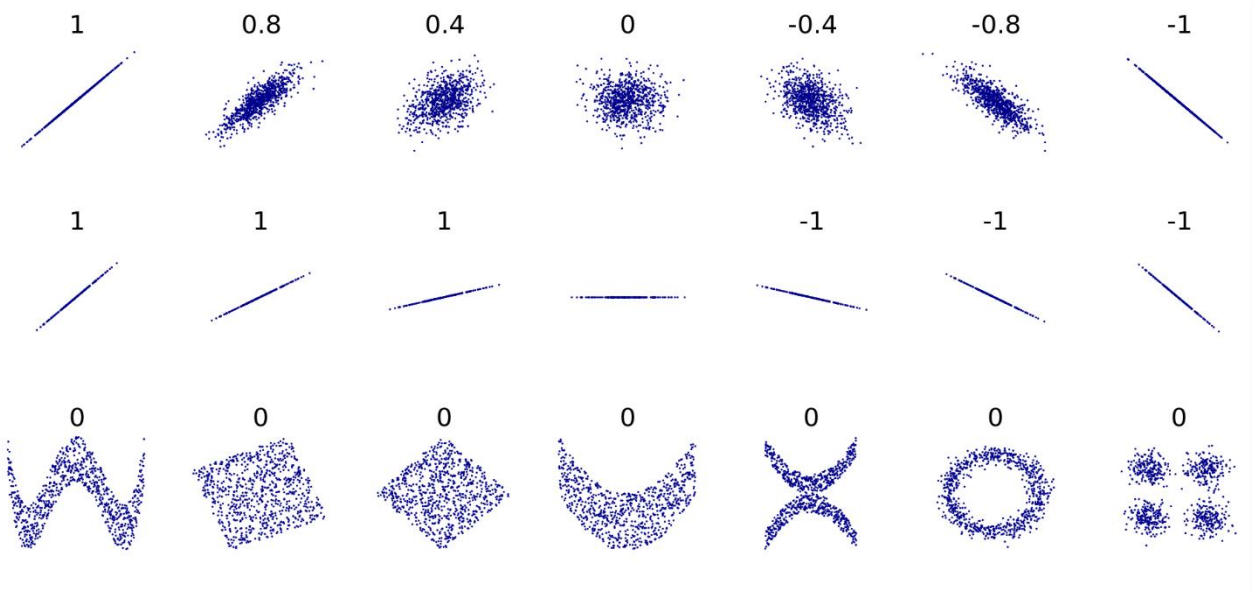

Figure 2: Examples of different bivariate datasets and their Pearson's correlation coefficient (©Wikimedia Commons)

The **covariance** is a measure for how much two variables change in the same direction. You can see from Equation (1) how this works: If the x value of a datapoint and the y value of a datapoint are both higher or both lower than the mean values of x and y, the datapoint contributes to a positive covariance (see red datapoints in Figure 3). However, if a datapoint has an x value that is lower than the mean x value and an y value that is higher than the mean y value (or vice versa), the datapoint contributes to a negative covariance (see blue datapoints in Figure 3). In a dataset, we can add up all those covariance contributions. If the total is positive, the Pearson's r will also be positive, and if the total covariance is negative, the dataset has a negative Pearson's r value.

We divide by the product of the standard deviations to normalize the Pearson's r index. If we would not do this, the Pearson's r will become very large for datasets in which variables can have very large values (such as geological ages in years) and very small for datasets with variables with very small numbers (such as concentrations of rare elements in seawater). Because we want our assessment of the correlation in the dataset to be independent of the unit we choose for our variables, we need to divide by the standard deviations.

Pearson's r is specifically defined to test **linear correlation** between variables, and it therefore not suitable to detect other patterns in bivariate data. You can see this by looking at the bottom row of Figure 2, in which the bivariate data clearly has a structure (and is therefore not *random*), but the Pearson's r is zero, which may cause you to interpret that the variables are totally unrelated. The same is true in Figure 4, which shows variables X and Y which have a perfect quadratic relationship, but the Pearson's r is almost zero. If you just looked at the Pearson's r in this bivariate dataset, you probably would have concluded that the two variables are unrelated, but this is not true. This shows you why it is always a good idea to plot your data and look at the data structure. Never rely on statistical tests alone!

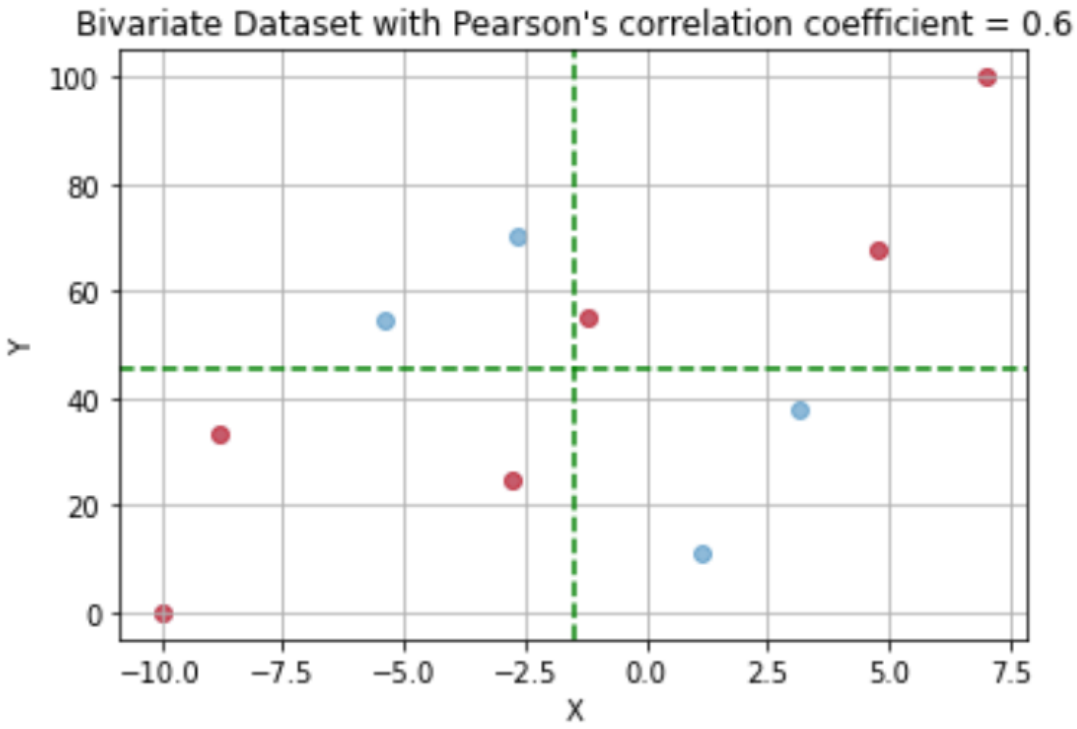

Figure 3: Scatterplot of a bivariate dataset with variables X and Y with a Pearson's r of 0.6. Dashed green lines highlight the mean values of X and Y. Datapoints in red have values for X and Y which either both exceed the mean value or are both smaller than the mean. These red points thus contribute to a positive covariance. The blue datapoints, on the other hand, have either higher-than-average X values and lower-than-average Y values, or vice versa. These blue points contribute negatively to the covariance. Since there are more red points than blue points and the red points are on average further away from the averages, the overall covariance is positive, and so is the Pearson's r.

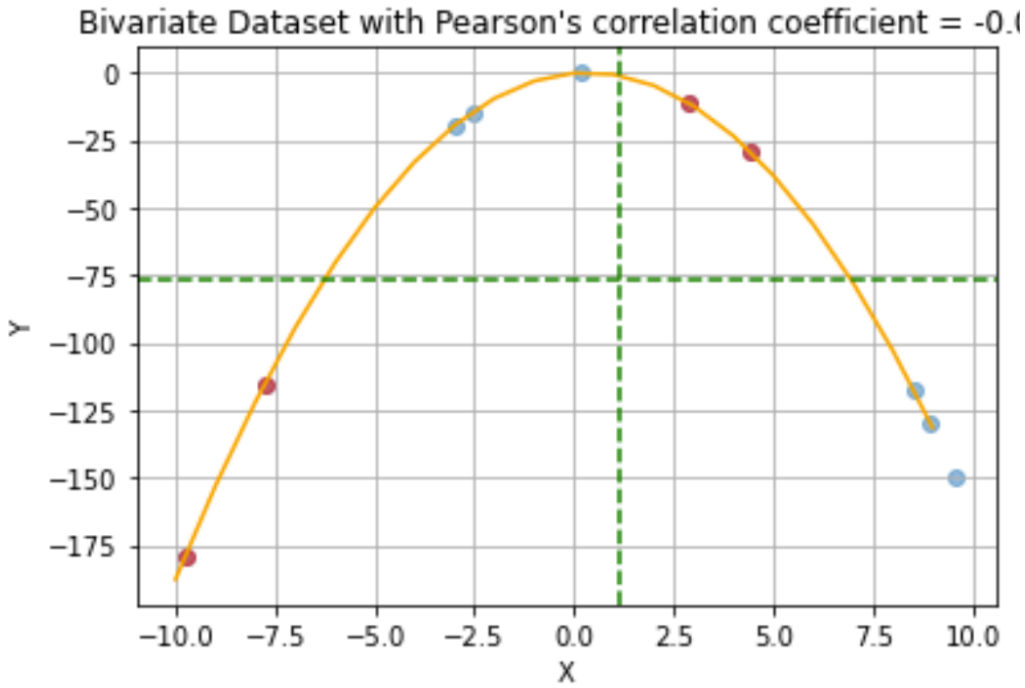

Figure 4: Scatterplot of a bivariate dataset with variables X and Y in which Y is related to X through a quadratic equation (of the form $y = a * x^2 + b * x + c$; see yellow curve). Dashed green lines highlight the mean values of X and Y. Datapoints in red have values for X and Y which either both exceed the mean value or are both smaller than the mean. These red points thus contribute to a positive covariance. The blue datapoints, on the other hand, have either higher-

than-average X values and lower-than-average Y values, or vice versa. These blue points contribute negatively to the covariance. Since there are about as many red points as blue points and they are roughly equally far away from the averages, the overall covariance is almost zero, and so is the Pearson's r.

II.4 Take Home Messages

- If two variables are *correlated*, they vary in the same direction
- We test *linear correlation* using the Pearson's correlation coefficient
- Correlation does not always mean *causation*. There can be *confounding variables* or the correlation may be coincidental (random)
- Never blindly trust a Pearson's r value (or any other statistical result), always interpret your data!

II.5 Extra reading

In this course, we only deal with Pearson's correlation coefficient. You have already seen that this coefficient is not ideal for each situation. In *Figure 4* you saw that it cannot be used to detect non-linear correlations. Another limitation of the Pearson's r is that it assumes that the distribution of the data is approximately normal, which is not always the case. There are other correlation coefficients that work better in scenarios when Pearson's r fails. You can read more about them here.

III. SIMPLE LINEAR REGRESSION

III.1 What is regression?

We have seen that correlation can tell us whether two variables co-vary in a linear way. However, that does not tell us anything about the shape of the *relationship* between the two variables. Regression is a statistical method used to test a relation between two variables. This relation is expressed as a formula of a line or eventually a plane or multi-dimensional shape in the case of more than two variables.

In most regression problems, there are ***independent* and *dependent* variables**. The independent variables have been measured more or less exactly, e.g. the distance along a transect measured using a measuring tape. We sometimes refer to the independent variable as the ***predictor variable*** because it is used to predict the other variable. In scatterplots, the independent variable is usually plotted on the horizontal axis (x axis).

In the formula which we try to find, the dependent variable(s) depend on the independent ones, in the sense that the formula *estimates* the value of the dependent variables from some value of the independent ones. We sometimes refer to the dependent variable as the ***response variable*** because it responds in some way to the predictor (or independent) variable. It is often plotted on the vertical axis (y axis). In most cases, the dependent variables are assumed to be subject to measurement error or uncertainty.

An example: you have taken soil samples along a sloping transect. The samples have been analyzed for organic matter content. The terrain elevation of each sample location is also known. You can check out the data in Table 1.

Table 1: Data on topsoil organic matter on a downslope transect.

Elevation (m)	Soil organic matter (weight %)
15	8.2
14.5	8.3
13.8	8.9
12.5	10.1
12.3	18.3
10.1	17.9
9.5	22.5
8.4	28.6
7.5	29.1
7.1	35

A useful starting point in your data analysis is drawing a graph of the data, since our mind is more sensitive to pictures than to numbers. When you have two variables that may be related somehow, it is always useful to make a scatterplot. In a scatterplot, the value of each variable is taken as an x and an y coordinate of a point in space. Each observation is plotted as a point in space (Figure 5).

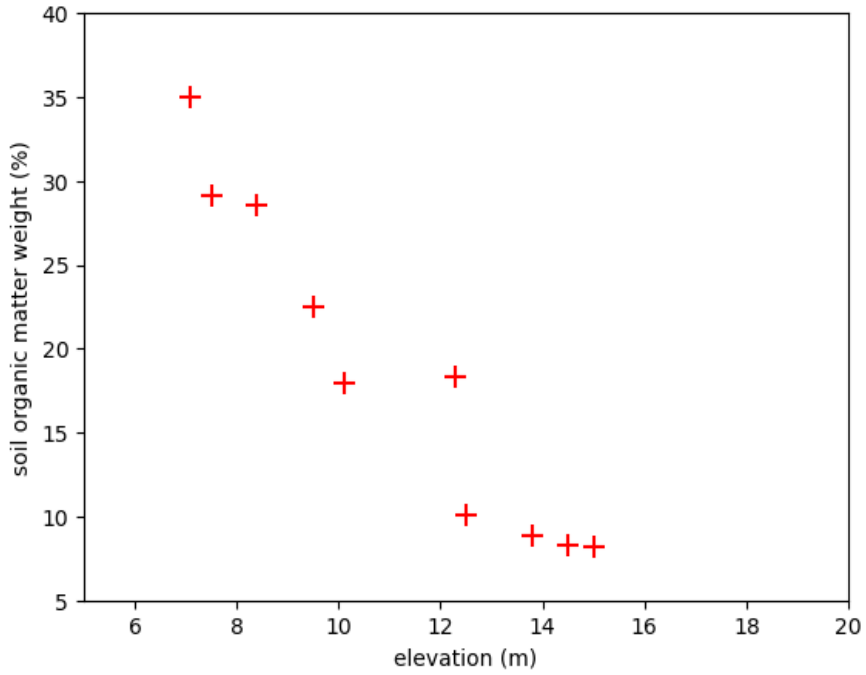

Figure 5: Scatterplot of the elevation and soil organic matter weight % variables.

A scatterplot shows a scatter of points. If you take elevation as the horizontal axis and soil organic matter content as the y-axis, a clear pattern emerges. Soil organic matter goes down as the elevation goes up. There may be several processes causing this phenomenon. For instance, soil erosion may have stripped organic topsoil from the higher places (try to think of other causes!).

At this point it is useful to think of the way you make your plot. We have to distinguish between *dependent* and *independent* variables. This also depends on what you know about the processes

behind the data, sometimes you do not know enough to make this distinction. This means that, even before you start to do data analysis, you have to think about potential causal relationships in your dataset! In this case, it is not very likely that the soil organic matter content has caused elevation differences of several meters. It is more likely that the soil organic matter depends on the elevation. As mentioned above, it is common practice use the horizontal (x-)axis for the independent variable and the vertical (y-)axis for the dependent variable.

Of course, you can describe this relation between organic matter and elevation using a correlation coefficient. It will result in a high correlation, in this case the Pearson's r is -0.96, which is very high. Remember that the further the Pearson's r is from zero, both towards 1 or towards -1, the stronger the linear correlation. But we can do better, for instance by describing *how fast* the organic matter content decreases with increasing elevation. We can try to write the relation between the two variables as an equation. This has advantages: With an equation you may be able to estimate organic matter content for points where you did not make an (expensive) analysis but where you know the elevation. An equation may also tell you more about the processes behind your observations.

As you can see from the Figure 6, the relationship between soil organic matter and elevation could be described by a simple linear equation.

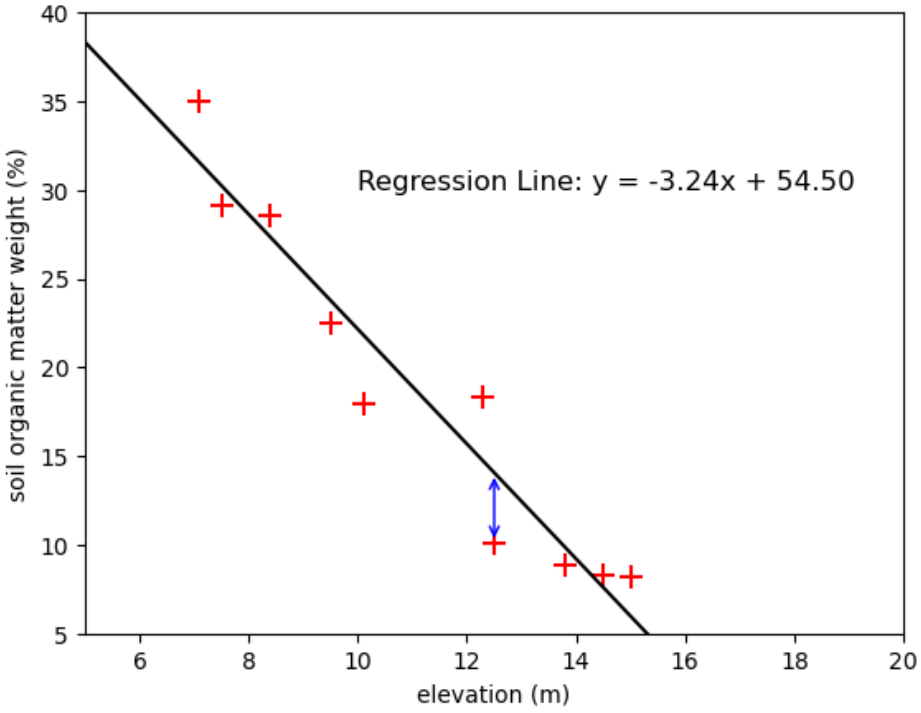

Figure 6: A possible equation for describing the relation between soil organic matter and elevation, and the deviation from the observed soil organic matter from the estimate according to this equation. The blue arrow indicates the size of the residual for one of these datapoints.

$$[Percentage\ soil\ organic\ matter] = [some\ constant] + [a\ second\ constant] * [elevation] \quad (2)$$

If we refer to our estimate for the percentage of soil organic matter as $\hat{y}$ (where the ^ stands for estimate, in contrast to the observed values y_i), and the elevation as x (with x_i referring to one observation of the variable x), we can write the equation as follows:

$$\hat{y}_i = a + b * x_i \quad (3)$$

Again, *x* (elevation) is the independent variable, from which the dependent variable, *y* (soil organic matter) is estimated. We call a statistical model like this a **simple linear regression**. It is "simple" because it only concerns one dependent and one independent variable. It is "linear" because the relationship between these two variables is described by a linear function (a straight line).

III.2 How to find the right line?

Of course, the line in Figure 6 can be drawn in several ways. Therefore, we need to make some choice which line is best, or what the best values of b_0 and b_1 are. To make that decision, we need a criterion. A logical choice is to find a line, such that all the deviations of the observed values of the dependent variables (y_i) from the line are as small as possible (see Figure 6). To do that, we can minimize the **sum of squares** of all the differences between the y values we measured (y_i) and the y values the line predicts ($\hat{y}_i$):

$$\sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \qquad (4)$$

In this sum, the $\hat{y}_i$ are the values of y estimated using the equation, and the y_i are the values observed at every point x_i. The summation is made for all *n* observed values of *y*. The deviations of the observed y_i from the regression line ($\hat{y}_i - y_i$) are usually termed *residuals*. The procedure of minimizing the sum of squared residuals is commonly called **Ordinary Least Squares Regression** (or OLS regression).

The explanation for why we are minimizing the sum of squares of the residuals goes beyond the scope of this course and has to do with the fact that this algorithm is the most efficient way to estimate the parameters of the regression line (*a* and *b*) following the *maximum likelihood theorem*. If you want to dive into the specifics, this is a good starting point. For the purpose of this course, you can remember that minimizing the sum of squares has the following neat benefits for our regression:

- It makes sure that all the deviations of points from the regression line are positive, so negative residuals also contribute to the total sum
- It penalizes points that are farther from the regression line extra strongly, because the square function augments the effect of larger residuals
- The sum of squared residuals is a measure of the amount of *variance* in the dataset that is not explained by the regression, and therefore has as specific statistical meaning. We will come back to this point later.

The derivation of the minimization of the sum of squares requires differential calculus and will not be given here. If you are interested in how the calculation for OLS regression works with an example, this is a good place to start. The result is a set of two equations (Equation (5 and (6) with two unknowns (the values of *a* and *b*) and coefficients calculated from the original observations x_i and y_i:

$$\sum_{i=1}^{n} y_i = \mathrm{a} * n + \mathrm{b} \sum_{i=1}^{n} x_i \qquad (5)$$

$$\sum_{i=1}^{n} x_i y_i = \mathrm{a} \sum_{i=1}^{n} x_i + \mathrm{b} \sum_{i=1}^{n} x_i^2 \qquad (6)$$

Here, n is again the number of observed values. The values of *a* and *b* obtained from these equations have the desired property of minimizing the difference between observed and estimated values for *y.* Rewriting, we obtain expressions for *a* and *b*:

$$b = \frac{\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} = \frac{SP_{xy}}{SS_x} \quad (7)$$

and

$$a = \frac{\sum_{i=1}^{n} y_i}{n} - b * \frac{\sum_{i=1}^{n} x_i}{n} = \bar{Y} - b\bar{X} \quad (8)$$

These equations look complicated but are in fact not as unfamiliar as you might think. The denominator in the equation for *b* is the same as the corrected sum of the products between all x_i and y_i (SP_{xy}). The numerator is the corrected sum of the squares of all x_i (SS_x). The corrected sum of squares is also used in computation of the variance (which is discussed in the Statistics part of this course). In the equation for *b*, the averages of the *y*'s and *x*'s, $\bar{Y}$ and $\bar{X}$ are used. The equation for *a* simply uses the formula for the regression line (Equation (3) and fills in the value we just calculated for *b* as well as the averages of the *x* and *y* values in the dataset ($\bar{X}$ and $\bar{Y}$). If we calculate the values for *a* and *b* in this way, we get *a* = 54.5 and *b* = -3.24 (see Figure 6).

Note that for this course, you do not have to be able to do the linear regression by hand (we have software like Python for that!). The derivation above serves merely to help you understand how linear regression works.

III.3 Does the regression line tell us anything meaningful about the data?

Having found an equation of a regression line does not make a real statistician happy yet. Using the algorithm of the previous section, you can find a regression line for any combination of variables. However, that does not mean that the regression line is meaningful. For instance, it could be that the slope parameter *b* does not deviate significantly from zero. If the slope of your regression line is zero, you have created a regression which is just a horizontal line, running through the mean value of all *y* values. That means that *y* does not depend on *x* at all! In such a case, simply using the mean value for *y* is just as good a predictor for unknown values of y as the regression equation, so the value of *x* does not predict anything about the value of *y* and your regression is meaningless.

If you find a regression line with a slope close to zero degrees, there is a chance that the regression line you drew through your data has this slope by coincidence. If you want to convince your fellow researchers that this regression is actually meaningful for predicting values of *y*, you have to prove that the slope you found is **significantly** different from zero. In other words: You would have to prove that it is very unlikely that your line has a non-zero slope just because your dataset is very noisy or scattered. The more the data points deviate from the regression line, the larger the uncertainty in your regression, so the size of your residuals tells you something about the uncertainty in your regression model.

For instance, in Figure 7 we see a dataset similar to that in Figure 6, but now the relation between the *x*'s and the *y*'s is less obvious. We can calculate a regression line using the OLS procedure

explained in the previous section. This results in the equation:

$$y = 10.29 + 0.099 * x \quad (9)$$

However, the regression line hardly differs from a horizontal line through the average of *y*, and the deviations of the observed *y*'s with respect to the line are very large compared to the slope of the line. In a case like this, the regression does not seem to be very useful for predicting *y* from *x*; the average of *y* would probably be just as good an estimate of any value of *y* than the y values the regression gives us.

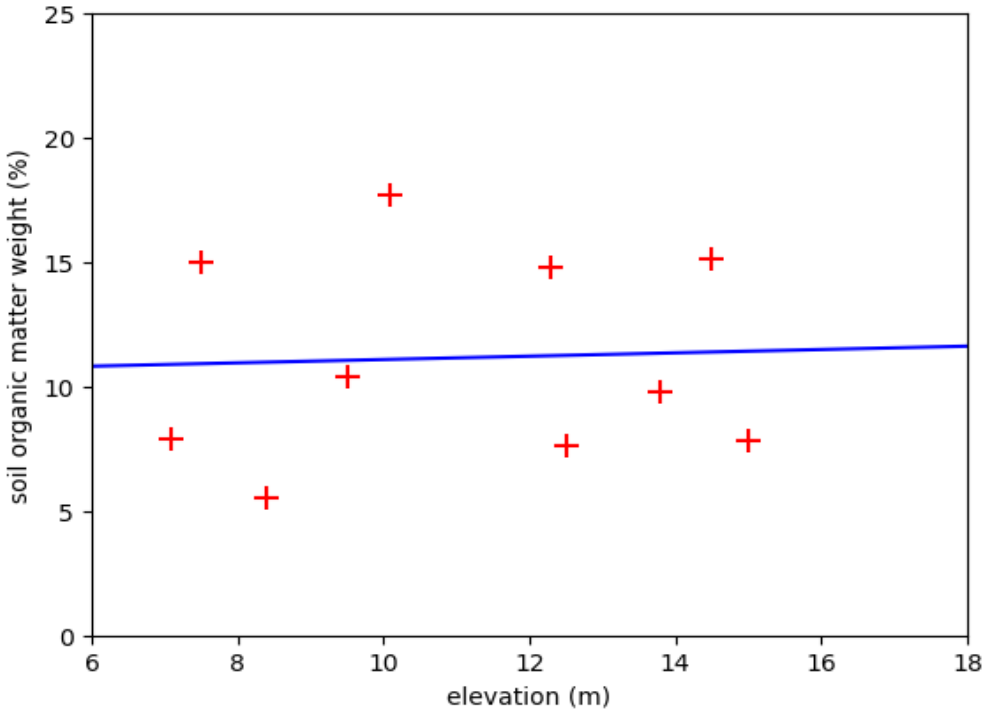

Figure 7: A near-horizontal regression line with large deviations may indicate that a relationship between the dependent and independent variable is absent.

Based on our intuition, we do not have much trust in the regression in Figure 7, but how can we prove whether your skepticism is justified? We can base our judgement of how useful a regression line is on comparing variances. In our regression problem, there are three sources of variation, each with their own variance.

1. The **total variance** of the original data, or the difference between y_i with respect to the mean y value ($y_i - \bar{y}_i$; the red lines in Figure 8)
2. The **variance of the residuals**, or de difference between the original data and the estimates ($y_i - \hat{y}_i$; the length of the blue lines in Figure 8).
3. The **variance of the *y* estimates** (or "variance of the regression") defined by the difference between the regression and the mean y value ($\hat{y}_i - \bar{y}_i$; the green lines in Figure 8)

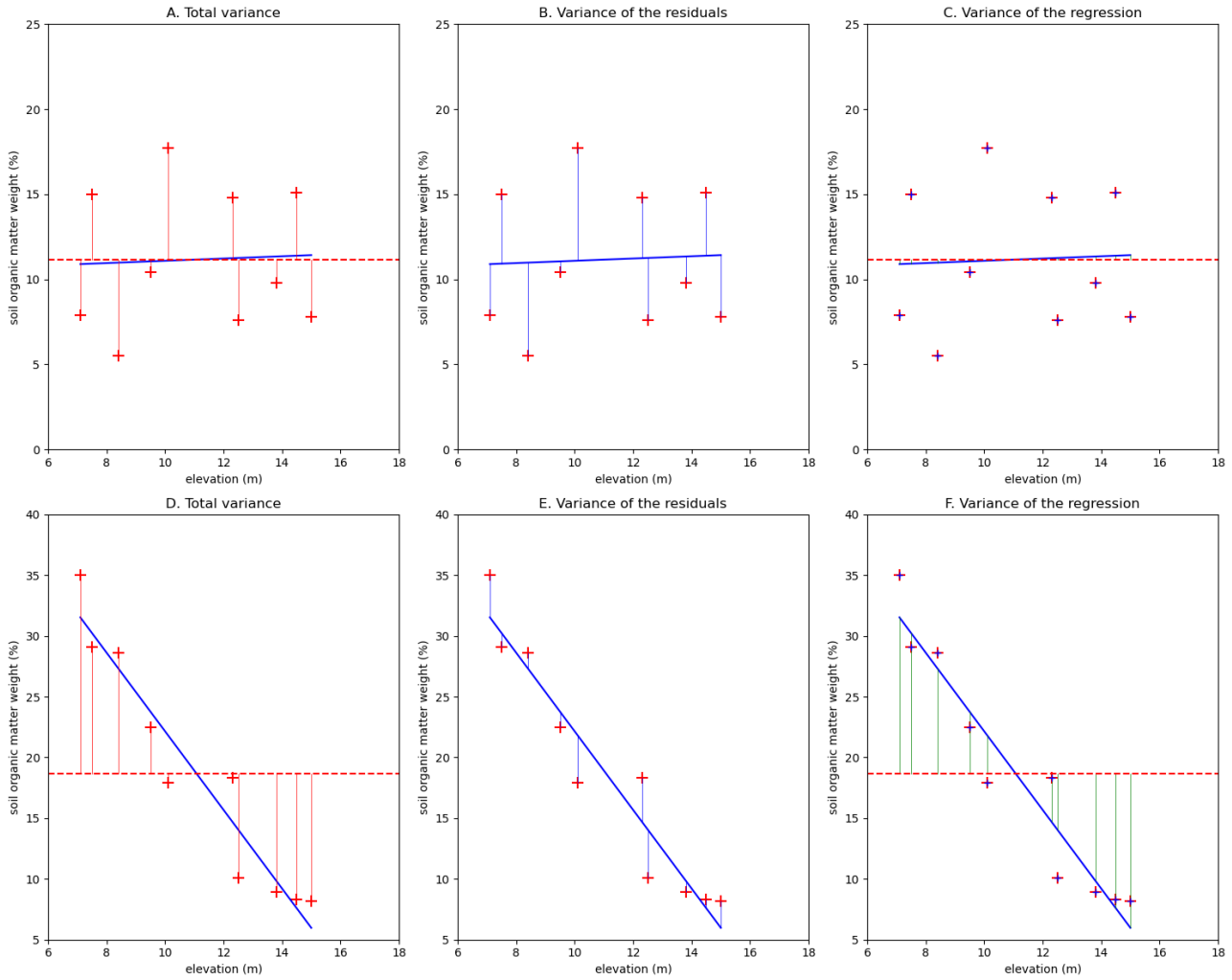

Figure 8: Regression lines and data points of the datasets in figures 5-7. Red crosses: the original data; Red lines in figures A and D: differences between the mean y value and the measured y values (total variance); Blue lines in figures B and E: Differences (residuals) between the original data and the y values estimated by the regression (Variance of the residuals); Green lines in figures C and F: Differences between the y values estimated by the regression and the mean y values (variance of the regression).

After calculating the parameters of a regression line, we can quantify these three sources of variance by calculating three sums of squares (all values squared and summed; see Equation 4). These sums form the basis for calculating variances:

1. The sum of squares of the original (observed) *y*'s, which represents the total variation in the data, denoted by SS_{total} or SS_T (this is the sum of the squares of the lengths of the red lines in Figure 8)
2. The sum of squares of the estimated *y*'s, the$\hat{y}_i$, or sum of squares of the regression, denoted by SS_R (this is the sum of the squares of the lengths of the green lines in Figure 8)
3. The sum of squares of the residuals ($y_i - \hat{y}_i$), or error sum of squares, denoted by SS_E (this is the sum of the squares of the lengths of the blue lines in Figure 8)

For these sums of squares holds:

$$SS_E = SS_T - SS_R$$

(10)

SS_T is the same as the sample variance of *y*:

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$$

(11)

SS_R, the sum of squares of the regression, is defined by:

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2$$

(12)

III.4 Goodness-of-fit

In a good regression line, the variance of the estimates $\hat{y}_i$ should be large compared to the total variance of the original y_i. Ideally, both variances should be equal. In that case there are no residuals, and all data points are exactly on the regression line. The quantity **goodness-of-fit**, denoted by R^2 ('**R-squared**') is a measure of how good the regression line fits the data. The *goodness-of-fit* is defined as the ratio between the sum of squares of the regression and the total sum of squares:

$$R^2 = \frac{SS_R}{SS_T}$$

(13)

In an ideal case, SS_R and SS_T are equal (and $SS_E = 0$; see Equation 10), so R^2 equals 1. In case the regression is very bad, SS_R is much smaller than SS_T and R^2 approaches 0 (and SS_E is large). So the higher R^2, the better the fit of the regression line. The square root of the R^2 value is equal to the absolute value of the Pearson's correlation coefficient, which we encountered in the II. CORRELATION section.

III.5 Testing the significance of a simple linear regression

Next, it should be shown that the *y*'s truly depend on the *x*'s, and that the regression line is a good predictor of *y*. When we do a regression, the theoretical population model for the regression line is:

$$\mathrm{y}_i = a + b * \mathrm{x}_i + \varepsilon_i$$

(14)

Here, *a* and *b* are the true population values of the coefficients of the regression line, and ε_i a random error of which the mean for all points *i* is zero and which is independent of the values of y_i. Ideally, the deviations of the data points from our fitted regression line ($\boldsymbol{y_i - \hat{y}_i}$) should completely represent these errors. In reality, our fitted regression line is always imperfect, so the values for *a* and *b* we obtain with our OLS regression will be different from the true *a* and *b* values of the population, which represent the true relationship between *x* and *y*. Because we do not know these true values, we cannot be completely certain how good our regression performs, but we can make a good estimate based on what we know.

Of course, when the points (y_i) are very close to the regression line ($\hat{y}_i$), there is much less "room" for the true relationship to be very different from our estimated line. In that case, the values for ε_i will be closely approximated by the deviations between our points and the regression line. To test this, we can show whether the regression line predicts the values of *y* from *x* better than just the mean of all *y*'s. This is the case when the variance of the residuals ($y_i - \hat{y}_i$) is small relative to the variance of the estimates ($\hat{y}_i - \bar{y}_i$). The smaller the variance of the estimates, the closer the slope

of the regression line is to zero, and the smaller the variance of the deviations should be for the regression to be meaningful. Therefore, the variance of the residuals is compared with the variance of the *y* values estimated by the regression line $\hat{y}_i$. The significance test which tells you if differences between these variances is large enough to conclude that the slope of the regression did not arise by chance is the **F-test**, also known as variance-ratio test.

The F-test is part of a statistical procedure called an **Analysis of variance (ANOVA)**. In an ANOVA, we aim to attribute parts of the variance in a dataset to different sources. In the case of a simple linear regression there are two sources: The regression and the residuals (or "noise" around the regression). The variance of those two sources adds up to the total variance. The F-test is outlined in the analysis of variance (ANOVA) table below.

Table 2: Analysis of variance table for a simple linear regression

Source of variation	***Sum of squares***	***Degrees of freedom (df)***	***Mean squares***	***F-test***
Regression	SS_R	*1*	$MS_R=SS_R/df$	MS_R/MS_E
Deviation (residuals)	SS_E	*n-2*	$MS_E=SS_E/df$	
Total variation	SS_T	*n-1*		

The sums of squares (SS_T, SS_R and SS_E) defined above must be converted into variances (or 'Mean squares', denoted by *MS*) by dividing them by the appropriate number of **degrees of freedom** on which they are based. The reasoning by which these degrees of freedom are derived is as follows:

- The total variance is the same as the sample variance (which you learned about in the Statistics part of this course) which based on *n* independent observations minus one for the estimate of the mean which you need to calculate the variance. You can say that calculating the mean value "costs" one degree of freedom.
- The mean squares of the regression (MS_R) is based on two "observations", namely the two coefficients of the regression equation (*a* and *b*). Therefore, the MS_R should require two degrees of freedom. However, every simple linear regression line passes through the mean value of the independent and dependent variables. Since we have already "spent" the degree of freedom associated with the mean when calculating the total variance (see above), we only need one extra degree of freedom to obtain the MS_R. This results in 2-1 = 1 degree of freedom.
- For MS_E, *n-2 = (n-1)-1* degrees of freedom are left over, because two degrees of freedom were used for the total variance and regression variance.

Finally, the F test statistic is calculated by MS_R/MS_E, so F is the ratio between the amount of variance explained by the regression and the amount of variance unexplained by the regression (note the difference between the F value and R^2 value! Compare with equation(13).

We test the significance with a one-sided interval, looking for the probability that F exceeds the critical value. This critical value depends on the number of degrees of freedom (*n-2*) and the threshold probability (α) we use to determine when we consider a result significant. Usually we take α = 0.05, which means that when our F-value exceeds the critical value, we have a 5% chance that

the regression result happens by chance. In that case, we are *95% confident* that the regression is statistically significant. If you want to learn more about how the F-statistic works, you can start by watching this explainer.

With the F-test we test the following hypothesis:

H_0: MS_R is not larger than MS_E

This is equivalent to saying that the scatter (variance) of the data points around the regression line (MS_E) is similar or larger than the variance of y_i (MS_R). In other words, if H_0 is true, you cannot conclude on a relation between *x* and *y*.

The alternative hypothesis is:

H_1: MS_R is larger than MS_E

This is equivalent to saying that the variance of the data points around the regression line (MS_E) is considerably *smaller* than the variance of y_i (MS_R). That means that the F value (MS_R/MS_E) must be too *large* to arise by chance. This is also equivalent to saying that the regression line *explains* a significant part of the variance in the dependent variable y_i. In normal language: The regression represents a relation between *x* and *y*.

In various textbooks we also see the hypothesis formulated in terms of the slope coefficient of the regression line: $H_0: b = 0$, $H_1: b \neq 0$. However, this is not the same as the previous hypothesis and holds the risk of drawing the wrong conclusions from the test. H_0 in this case could be taken as evidence that there is a zero slope angle, while in reality the F test does not allow you to draw conclusions on the angle of the line (or the strength of the relationship). In general, in most of these cases *b* would indeed be near-zero (which is a consequence of the absence of a relation), but it can also attain higher or lower values. This is also shown by generally high values for the confidence interval for *b*. Below, we will work out an example to show this.

Consider for instance Figure 6. It is obvious that in Figure 6 the MS_E and MS_R will differ much. On the other hand, in a situation like that of Figure 7, the difference between MS_E and MS_R is quite small. Here is a worked example based on Figure 6:

Table 3: Example of calculations for the total sum of squares and the sum of squares of a simple linear regression

Elevation (x)	***Soil organic matter % (y_i)***	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
15	8.2	-10.5	110.0	5.9	-12.7	161.6
14.5	8.3	-10.4	108.0	7.5	-11.1	123.1
13.8	8.9	-9.8	95.8	9.8	-8.8	78.0
12.5	10.1	-8.6	73.8	14.0	-4.6	21.4
12.3	18.3	-0.4	0.15	14.7	-4.0	15.8
10.1	17.9	-0.8	0.62	21.8	3.1	9.9
9.5	22.5	3.8	14.5	23.7	5.1	25.8
8.4	28.6	9.9	98.2	27.3	8.6	74.6
7.5	29.1	10.4	108.4	30.2	11.6	133.4
7.1	35.0	16.3	266.0	31.5	12.8	165.0

Elevation (x)	***Soil organic matter % (y_i)***	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
n = 10	$\bar{y} = 18.7$		∑=SS_T=875.5			∑=SS_R=808.6

The goodness-of-fit $R^2 = SS_R/SS_T$ becomes 808.6 / 875.5 = 0.92. Working out the analysis of variance (ANOVA) table we get for F:

Table 4: Analysis of Variance (ANOVA) table for a simple linear regression

Source of variation	***Sum of squares***	***Degrees of freedom (df)***	***Mean squares***	***F-test***
Regression	$SS_R = 808.6$	*1*	$MS_R = SS_R/df = 808.6/1 = 808.6$	$MS_R/MS_E = 808.6/8.4 = 96.65$
Deviation / residuals	$SS_E = SS_T - SS_R = 875.5 - 808.6 = 66.9$	$n-2 = 10-2 = 8$	$MS_E = SS_E/df = 66.9/8 = 8.4$	Critical value F at 1% significance and $df_1=1$ $df_2=8$: 11.26
Total variation	$SS_T=875.5$	$n-1=9$		

From calculations in the table above, the value of F becomes 96.65. Looking up the critical value of F for 1 and 8 degrees of freedom, we get a value of 11.26. The value of F is far greater, so we can reject H_0.

If we do the same for the data on the left side of Figure 7, the results are SS_T = 155.8, SS_R = 0.8, SS_E = 155.0 and F = 0.8 / 155.0 = 0.005, with the same degrees of freedom. Clearly, the variance of the deviations of the data points from the regression line is far greater than the variance of the regression line itself, and is nearly equal to the total variance. The resulting F is far below the critical value, even if a lower significance level is taken. The regression line thus does not specify any significant relation between *x* and *y*.

III.6 How to proceed with a poorly fitting regression line

If the statistical tests above do not confirm a significant fit to the data, this is not the end of the regression analysis yet. It may help much to consider the causes of a poor fit and apply remedies for these causes:

Firstly, our straight-line equation may not be appropriate. Rather, a curved line may be a better approach of the relation between the dependent and independent variable. This may be guessed from the scatter plot of the data, for example in Figure 4. A similar dataset is plotted below in Figure 9 showing the linear regression on top of the points.

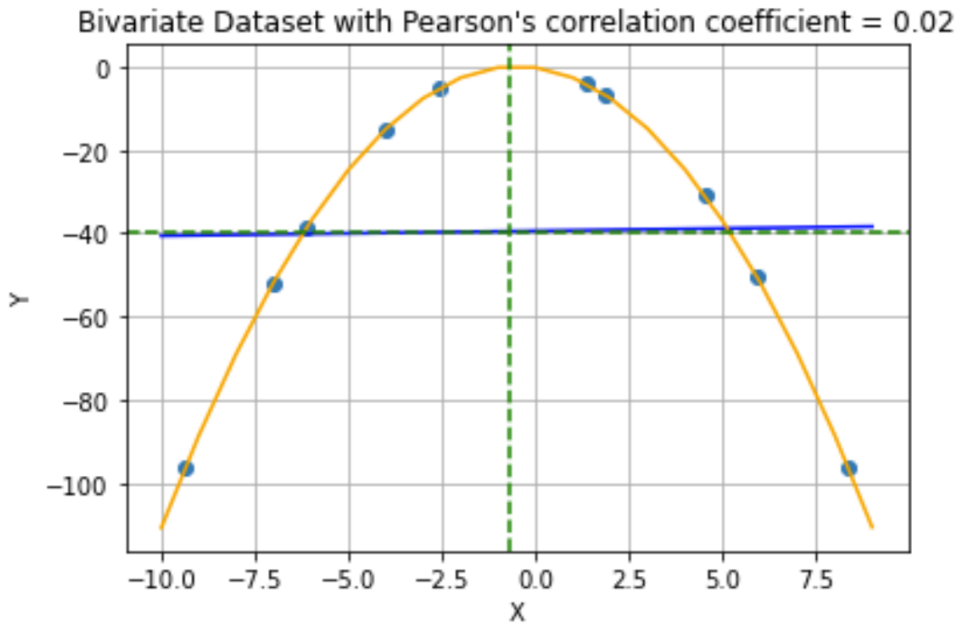

Figure 9: Scatterplot of a bivariate dataset with variables X and Y in which Y is related to X through a quadratic equation (of the form $y = a * x^2 + b * x + c$; see yellow curve). Dashed green lines highlight the mean values of X and Y. The blue line highlights the best fit of a linear regression, which has a very poor R^2 value. Note that the linear regression is almost on top of the horizontal line marking the average value for Y.

In a case like this, the residuals are rarely evenly scattered along the regression line: In one part (in this case for very negative or very positive x values), most of the points are situated below the line, in another part (in this case for x values close to zero) above. The remedy is performing curvilinear regression, by fitting a regression equation that results in a curved line, which is discussed in Section III.8 Extra reading: Calculating confidence intervals on the regression constants

Secondly, there may be more than one independent variable that determines the variation in the dependent variable. In that case, the regression may be significant, but still with a large variance of the residuals. If you have measurement data of these variables, you may try to fit a regression equation that contains more than one independent variable (multiple linear regression). In that case the equation represents a plane rather than a line. This technique is treated V. MULTIPLE REGRESSION.

Thirdly, the variance in the data may indeed be simply too high to discover a meaningful relation, as in the H_0 hypothesis of the F test described above. Even then it may be possible to improve the regression by taking a closer look at the data points that deviate most from the regression line (and the other data points). These points may be the result for instance of measurement errors. If you really have good reasons to assume that measurement errors are the cause (for example: if you know that that particular measurement did not go as planned), then you can choose to exclude these data from your analysis. However, **never exclude data without good reasons** and without stating the reasons why! In general, *outliers* in the data are suspect.

Outliers are data points that are far beyond the range of the other data. What defines whether a datapoint is an outlier can vary a lot per project and is inherently a subjective question. Some researchers use a threshold for determining outliers that is based on the standard deviation in the dataset. For example, any point that is more than 3 standard deviations from the mean value is an outlier). This is no foolproof way to detect outliers, because datapoints can also be further away from the mean by chance, or the outlier may represent a real phenomenon in the data that may be important to analyze. The question of which data represents an outlier is a great example of an

important lesson in statistics and data analysis:

Statistical tests are useful tools, but in the end, it is always the researcher who interprets the data and draws scientific conclusions. Statistics will not do the research for you, and all data analysis procedures require you to make subjective decisions.

In Figure 10A, three outliers strongly increase the variance of the residuals, and will result in a low significance of the F test. Removing these outliers (Figure 10B) results in a much stronger correlation coefficient and a different regression equation.

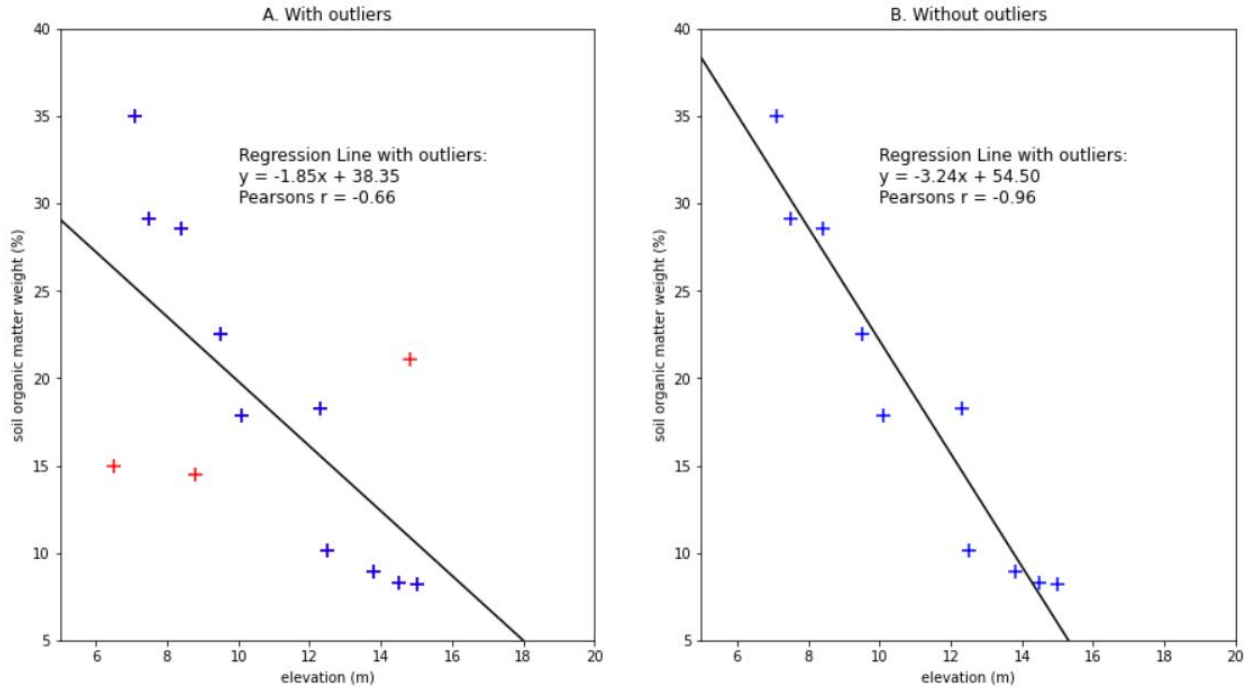

*Figure 10: Two linear regressions between soil elevation and soil organic matter. **A** (left) shows three points in red which could be considered outliers. In **B** the outliers are removed. Note how the correlation coefficient and the regression equation change when we remove the potential outliers.*

Outliers may not only result in a poor fit, but they also result in a severe distortion of the regression line, especially if situated on the boundaries of the range of the independent variable. For instance, in Figure 10, the data set with and without outliers would result in a regression line that passes the F test. However, the outliers cause a large difference of the slope of the line!

III.7 Take Home Messages

- We can test whether two variables are linearly related using a **simple linear regression**
- To find the best straight line fit through bivariate data, we use the **Ordinary Least Squares** method
- The **goodness of fit** (R^2) tells us how well the line approximates the datapoints. This tells us something about the **strength of the relationship**
- We use an **ANOVA** to determine whether the relationship is **significant**. This can be concluded based on the **p-value** using an **F test**
- **Outliers** can influence our conclusions about the relationship between two variables. The criteria with which to determine whether a datapoint is an outlier is subjective and depends on the research question and type of data
- Based on the mean squares of the unexplained variance, we can calculate the uncertainty on the slope and intercept of the regression.

III.8 Extra reading: Calculating confidence intervals on the regression constants

The coefficients of the regression line are generally calculated from a *sample* of a larger population. When you estimate a population mean and variance from a sample from that population, these statistics have an estimation error, depending on the size of the sample. The larger the sample, the closer the estimate will be to the real mean of the entire population. The same holds for regression coefficients derived from a sample. In some cases, it is necessary to know these estimation errors, in particular when you estimate physical quantities from a regression line.

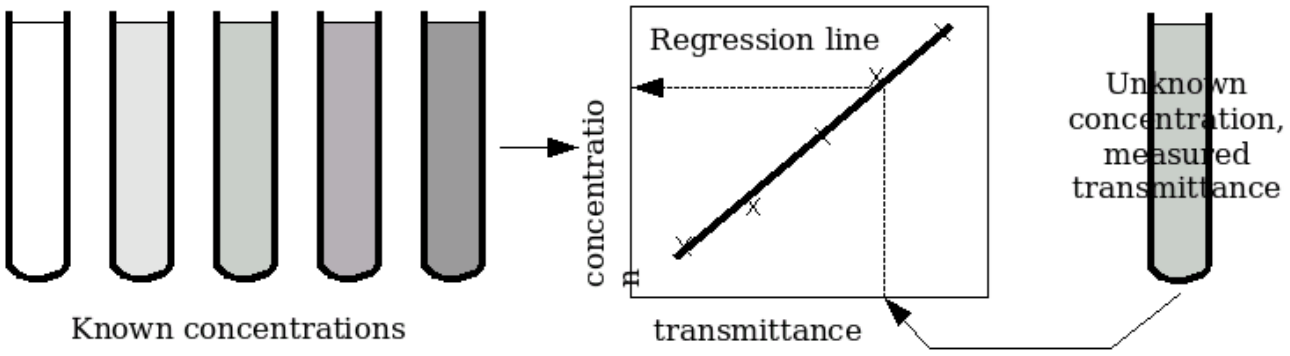

Figure 11: Determination of the concentration of a chemical compound in a solution using a colorimeter.

As an example, consider common laboratory practice. Say, you have a colorimeter, which measures light transmittance through a colored solution of some compound. The transmittance is linearly related to the concentration. The usual measurement procedure is to make a small series of solutions with increasingly higher, but known, concentrations of the compound (called a calibration set), and measure the transmittance with the colorimeter. Next, a regression line is calculated to find the relation between transmittance (independent variable) and concentration (dependent variable). After that, you can measure the transmittance of each unknown concentration, and calculate the concentration based on the regression line. Of course, there is some measurement error in each colorimetric measurement resulting in statistical estimation errors in the coefficients of the regression equation (*a* and *b*, see (3)). If you want to quantify this error, you need to know the estimation variance of these coefficients. For the slope coefficient *b* this is:

$$\sigma_b^2 = \frac{MS_E}{\sum(x_i - \bar{X})^2}$$

(15)

In this formula, MS_E is the mean squares of the deviations from the regression discussed above, and the term in the denominator represents the variance of the independent variable (*x*). We can define a confidence interval around *b* by using the Students' *t* distribution:

$$b_1 \pm t_{1-a/2,\, n-2}\sqrt{\sigma_b^2}$$

(16)

In this formula, *a* is the confidence interval (often 0.05, or 95%, in which case all possible values of *b* lie with 95% certainty within the interval), and *n-2* the degrees of freedom for *t*. For the variance of the intercept *a*, a similar formula holds:

$$\sigma_a^2 = MS_E \frac{\sum x_i^2}{n * \sum(x_i - \bar{X})^2}$$

(17)

Note the similarities and differences with (15). To calculate the confidence interval around *a*, we can use (16) and replace σ_b with σ_a. The estimation variance for the values of y_i as estimated from x_i with help of the regression line can be determined as:

$$\sigma_{\hat{y}_i}^2 = MS_E \left(\frac{1}{n} + \frac{(x_i - \bar{X})^2}{\sum(x_i - \bar{X})^2} \right)$$

(18)

Again, the specification of a confidence interval is similar to that of *b* (see (16)), so the confidence interval would be given as:

$$analysis\ result = estimated\ concentration \pm t_{critical\ value} * \sigma_{y_i}$$

(19)

In the example above, you would use this formula to express the statistical error in your analysis result. In any linear regression project, you can use this function to calculate a **confidence envelope** around your regression line for any *x* value.

These formulae are also useful to test if there is a significant time trend in data, for example in climate or river discharge data. In those cases, the amount of data is large, and the regression line usually has small slopes (*b* values). Remember that an ANOVA only tests to what extent the regression line can be used as a predictor of the dependent variable; it does not test whether the slope departs significantly from zero. By using (15), you can really test the presence of a linear trend.

IV. SIMPLE NON-LINEAR REGRESSION

IV.1 How to get a curved regression line - transformations.

Once you have decided that a curved line should represent the relation between the variables better than straight regression line, there are two basic ways to proceed. The first one is applying a transformation to one of the variables. For instance, take the logarithm of y, which results in a regression equation like this:

$$\log(y) = a + bx \qquad (20)$$

This results, after some algebra, in an **exponential relation** between x and y:

$$e^{\log y} = e^{a+bx}$$
$$y = e^{a}.e^{bx}$$
$$y = ce^{bx} \qquad (21)$$

Since e^a is a constant, it is replaced by the new constant *c* for convenience.

The **log-transformation** is a smart way to turn a (more complicated) exponential relationship into a (more simple) linear relationship. Of course, other transformations can be applied as well. The choice will depend on your theoretical knowledge of the processes behind the relation, or a relation you assume based on the scatter plot of the data. Since you turn the regression problem into a simple linear regression, further treatment of the regression analysis is the same as that for linear equations discussed above. However, be careful! The **constants** you fit (e.g. "slope" *b* and "intercept" *c* in Equations 19 and 20) will have a **different unit** and a **different relationship with the variables** (*x* and *y*) compared to the *a* and *b* in a non-transformed simple linear regression.

As an example, we will be looking at a dataset that contains information about the concentration of CO_2 in the atmosphere measured (in parts per million by volume, or ppmV) over the period 1850-2022. This dataset is obtained from the website Our World in Data[1], which is an excellent source of up-to-date information about climate, food, economic development, biodiversity, and other pressing societal issues.

If we plot the measured CO_2 concentrations against time (Figure 12), we can see that there is a positive relation between atmospheric CO_2 concentrations and time. This should not surprise us too much; we all know CO_2 concentrations have been increasing since the Industrial Revolution. A linear regression gives a significant relation: The F test value = 613.5 ($p << 0.01$), the goodness of fit R^2 = 0.824, and the regression equation is:

$$pCO_2 = -1008 + 0.686 * t \qquad (22)$$

Here, *t* is time in years and pCO_2 is the concentration (or "partial pressure", hence the "p") of CO_2 in the atmosphere.

[1] Our World in Data and Max Roser, "OWID Homepage," *Our World in Data*, March 25, 2024, https://ourworldindata.org.

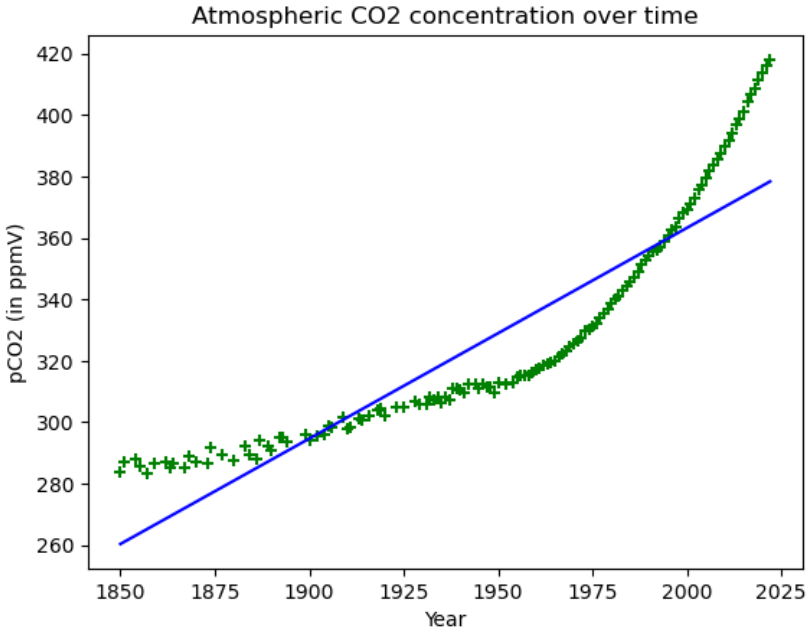

Figure 12: Plot of atmospheric CO_2 concentration measured over time (green dots). The blue line highlights the result of a simple linear regression.

However, in the plot in Figure 12 we also see that the relation me be not linear but rather a curved line. The increase in CO_2 concentration is also getting stronger as time progresses. This suggests an exponential relation rather than a linear one. We could test this hypothesis by taking the natural logarithm of the CO_2 concentration and testing its relationship over time. To make this easier, we first modify the variables a bit so time starts at 1850 and pCO_2 is expressed as a value relative to the pre-industrial concentration (~280 ppmV). This way, our exponential relationship moves through the origin (0,0) at a meaningful place. You see the result of this transformation and the linear regression in Figure 13.

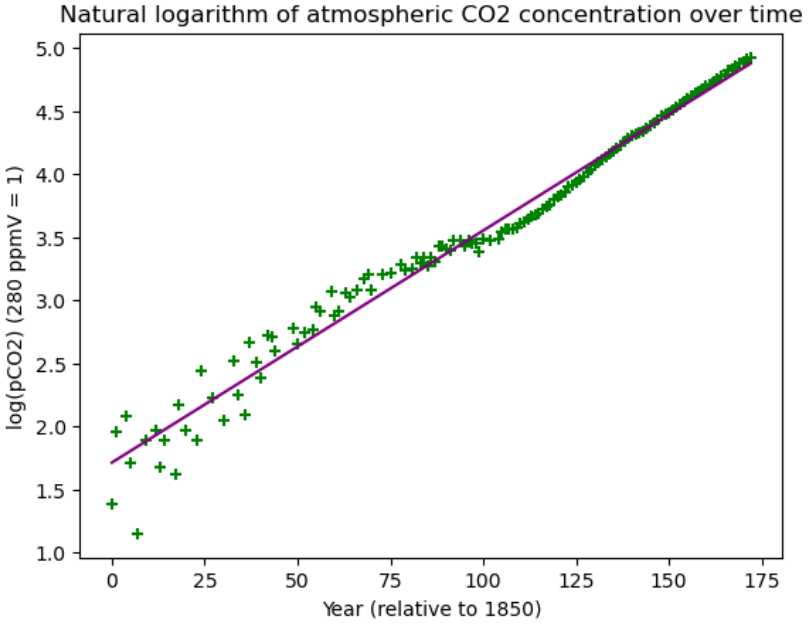

Figure 13: Plot of natural logarithm of atmospheric CO_2 concentration vs time (green dots). The pruple line highlights the result of a simple linear regression on the log-transformed data.

After taking the natural logarithm of the CO_2 concentration, we can compute a new regression line:

$$\log(pCO_{2_{rel}}) = 1.716 + 0.0184 * \mathrm{t}_{rel}$$

(23)

Here, t_{rel} is the time in years relative to 1850 and pCO_{2rel} is the CO_2 concentration relative to the reference value of 280 ppmV. The regression is significant again (F = 5689, N = 133, p << 0.01) and the goodness-of-fit has improved a lot: R^2 = 0.977. Removing the log CO_2 according to **Equation 20**, this results in the following exponential equation:

$$pCO_2 = \mathrm{e}^{1.716} * \mathrm{e}^{0.0184 * \mathrm{t}_{rel}} + 280 = 280 + 5.56 * \mathrm{e}^{0.0184*(\mathrm{t}-1850)}$$

(24)

This exponential regression line is shown in Figure 14, where the logarithms have been transformed back to linear values and the *pCO₂* and *t* values are calculated back to their original values.

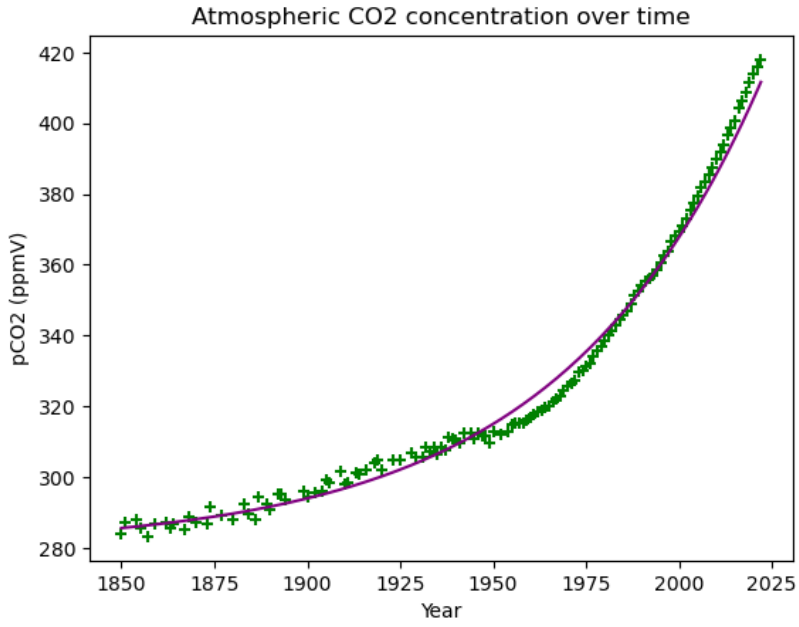

Figure 14: Plot of atmospheric CO_2 concentration measured over time (green dots). The purple line highlights the result of a simple linear regression on the log-transformed data after converting the data back to the linear scale.

IV.2 How to get a curved regression line - higher order polynomials.

The second approach to quantify the non-linear relationship between two variables is to use a higher order **polynomial** for the regression equation. A polynomial function is similar to the linear equation (see **Equation 3**), but it contains terms containing the independent variable (*x*) to a higher power (2, 3, 4, …). The regression equation for a polynomial looks like this:

$$Y = a + \mathrm{b} * X + \mathrm{c} * X^2$$

(25)

Equation 24 shows the formula for a **second order polynomial**, but higher order terms can be added indefinitely:

$$Y = a + \mathrm{b} * X + \mathrm{c} * X^2 + \mathrm{d} * X^3 + \cdots + b_m * X^m$$

(26)

The **order** of the polynomial is defined by the highest power to which the independent variable is raised (marked by the variable *m* in Equation 25). Following this logic, the simple linear relation discussed above is a **first order polynomial**, with the independent raised to the first power only. This

type of regression analysis is called **polynomial regression**. Typically, this is applied when there are no theoretical reasons or before-hand knowledge from which you may select another transformation.

If you have a 2nd order polynomial, you will have to find three coefficients for the regression equation (*a*, *b* and *c* in Equation 24). An *n*th order polynomial requires *n* + 1 coefficients: One for every power to which we raise the independent variable plus one for the intercept (*a*). In **IV.5 Extra reading: The mathematics behind a polynomial regression** I list some matrix algebra, to show you how polynomial regression is done. In practice, you never need to do this by hand, and (like all Extra Reading sections) this is not part of the material you need to know for the exam. All statistical software packages contain options for finding any regression equation. Do not be scared of the equations, you don't need to remember them, but it can help you to better understand how regression works if you try to grasp how they are formed.

IV.3 Judging the significance of a polynomial regression and the problem of overfitting

Just like with a simple linear regression, we use variance analysis to judge the significance of a polynomial regression line. However, we must do more than just one test: We have to also decide which polynomial order fits best: the first, second, third or *n*th order. In general, addition of an extra higher order term will result in a better fit of the regression equation to the data, as shown by an increasing goodness-of-fit, R^2. The higher the order of the polynomial equation, the more curves appear in the regression line. Ultimately, if *m* (order of the equation) equals *n-1* (*n* is the number of data points), the regression line exactly follows the data! Figure 15 shows an example. In this example, the R^2 value of the 5th order polynomial regression will be exactly 1.

This poses a problem, because the chance is high that this regression does not result in a meaningful regression equation. Remember that the goal of a regression is to find a relationship between two (or more) variables based on a limited sample of a large population of data (a sample of 6 in the case of Figure 15). If we force our regression to pass exactly through our datapoints, it is likely that the curve is not very good at estimating unknown data in the population. Furthermore, Figure 15 shows that higher order regression equations are highly sensitive to **outliers** in the data. See for instance the rightmost data point which causes the higher order regression lines to have very steep slopes. This is dangerous if we want to interpret the result of our regression and extrapolate it to the entire population!

Another way to think about this is that the general assumption for regression is that the original data satisfy a relation between the dependent and independent variable, plus a random error:

$$Y = f(X) + e \qquad (27)$$

No data is perfect, and all datapoints we use to do a regression will contain some error (*e*). It is very likely that the curves in the regression line created by the highest order terms are influenced by these random errors in the data. In other words: The higher our polynomial order, the more trust we have that our data is a flawless description of the relationship between the dependent and independent variable. The problem we encountered with the fourth and fifth order polynomial first in Figure 15 is called **overfitting**.

So how do we determine the optimal order for our polynomial equation? In Figure 15, the second order and third order equations closely resemble each other. This means that adding an extra term to the second order polynomial regression does not much improve the result. This is a good indication that the third and higher order terms may not be very meaningful to explain the relation

between *Y* and *X*.

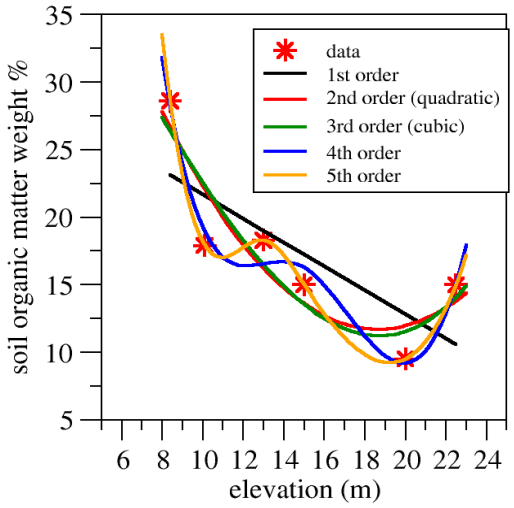

Figure 15: Example of fitting increasingly higher order polynomial curves through a dataset with 6 datapoints. Note that the 5th order polynomial (n-1st order) is a perfect fit, because it goes through every datapoint.

If we want to be sure about our choice of the right order of the polynomial, the significance of each regression fit should be tested. This can be done with the analysis-of-variance (ANOVA) test we used for simple linear regression. For this test, the ANOVA table must be expanded to also include the lower order regression lines and their deviations. For the simple first order (linear) regression line, we calculate the **sum of squares of the regression** (SS_{R1}) by subtracting the mean of the observed values of *Y* ($\bar{Y}$) from the estimates of *Y* for all points (*i*) according to the regression line ($\hat{y}_i$) and squaring and summing these:

$$SS_{R1} = \sum_{i=1}^{n} (\hat{y}_i - \bar{Y})^2 \qquad (28)$$

We do the same for the second order regression line, resulting in a sum of squares due to the second order regression, SS_{R2}, and repeat the same for the third and higher order regressions if necessary. For each regression line, we also compute the corresponding error sum of squares:

$$\begin{aligned} SS_{E1} &= SS_T - SS_{R1} \\ SS_{E2} &= SS_T - SS_{R2} \\ SS_{E3} &= SS_T - SS_{R3} \\ &\vdots \\ SS_{Em} &= SS_T - SS_{Rm} \end{aligned} \qquad (29)$$

The next step is then to find out whether the higher order regression line actually has *improved* the regression. Remember, a regression line is better when the variance contained in the regression line (below calculated as the mean squared deviation of the regression: MS_R) is larger with respect to variance of the deviations from the regression line (the mean squared error of the regression MS_E). If, for instance, the quadratic regression line (second order polynomial) is better than the first order regression line, the SS_{R2} should be significantly larger than the SS_{R1} (or SS_{E2} should be significantly

smaller than SS_{E1}). To test the significance of the contribution of the second order term, we therefore calculate the difference between SS_{R2} and SS_{R1}. This value has only one degree of freedom, so the corresponding mean squared error of this contribution, $MS_{R2\text{-}R3}$, is the same as SS_{R2}-SS_{R1} (because we divide by 1). Next, we can calculate an F-test value by dividing with the mean squared error of the second order regression line:

$$F_{adding\ a\ second\ order} = \frac{MS_{R2-R1}}{MS_{E2}} = \frac{SS_{R2} - SS_{R1}}{SS_T - SS_{R2}}$$

(30)

Below you find the complete ANOVA table for a 3rd order (cubic) regression:

Table 5: ANOVA table for a third order polynomial regression. Note that the fifth and sixth lines indicate the sum of squared deviations between a first and second order polynomial and between a second and third order polynomial, respectively.

Source of variation	***Sum of squares***	***Degrees of freedom (df)***	***Mean squares***	***F-test***
Linear regression (1st order)	SS_{R1}	*1*	$MS_{R1}=SS_{R1}/df$	MS_{R1}/MS_{E1}
Quadratic regression (2nd order)	SS_{R2}	*2*	MS_{R2}	MS_{R2}/MS_{E2}
Cubic regression (3d order)	SS_{R3}	*3*	MS_{R3}	MS_{R3}/MS_{E3}
Added to linear regression by quadratic regression	$SS_{R2-1}=SS_{R2} - SS_{R1}$	*1*	MS_{R2-R1}	MS_{R2-R1}/MS_{E2}
Added to quadratic regression by cubic regression	$SS_{R3-2}=SS_{R3} - SS_{R2}$	*1*	MS_{R3-R2}	MS_{R3-R2}/MS_{E3}
Deviations from cubic regression	SS_{E3}	*n-4*	$MS_{E3}=SS_{E3}/df$	
Total variation	SS_T	*n-1*		

Using this ANOVA, not only the significance of each individual regression line can be tested, but also the contribution by the successive additions of higher order terms. If for instance, the last F value in the table (that of the cubic X^3 term, sixth line from the top) is below the critical value ($p > 0.05$). From this we can conclude that it makes no sense to add a X^3 term to the equation. This is then automatically true for any higher order terms as well, because including higher order terms only reduces the degrees of freedom and the difference between the sums of squares. If the F value for the difference between polynomial orders is above the critical value ($p < 0.05$), it may be useful to add a fourth order term. In natural datasets, this rarely occurs because in most cases a second or third order polynomial is sufficient to explain the data.

IV.4 Take Home Messages

- Sometimes, a non-linear regression can be converted to a simple linear regression by **transforming** the variables. A common example of this is the **logarithmic transformation**.
- Polynomials are a family of curved lines described by functions with increasingly larger numbers of terms in which the independent variable is raised to increasingly higher powers. The highest power to which the independent variable is raised determines the **order** of the polynomial equation.
- Higher order polynomial fits achieve ever higher goodness-of-fit statistics (R^2). Fitting (very) high order polynomials to limited data increases the risk of **overfitting**.
- An ANOVA can be used to test whether a higher order significantly enhances the quality of the regression and may help to prevent overfitting.

IV.5 Extra reading: The mathematics behind a polynomial regression

Let's look at the calculations underlying the fit of a polynomial regression. By analogy with the equations we use to do a simple linear regression, we need to solve *m* simultaneous equations derived from the data to find the optimal fit for a polynomial regression. Equations 5 & 6 for a simple linear (1st order polynomial) are repeated here:

$$\sum_{i=1}^{n} y_i = \mathrm{a} * n + \mathrm{b} \sum_{i=1}^{n} x_i \qquad (31)$$

$$\sum_{i=1}^{n} x_i y_i = \mathrm{a} \sum_{i=1}^{n} x_i + \mathrm{b} \sum_{i=1}^{n} x_i^2 \qquad (32)$$

Again, we are not deriving these formulas here, but if you are interested to find out how this works "under the hood", you can have a look here. If not, you may assume these functions are correct. If we generalize these equations to calculate coefficients for higher order polynomials, we get the following family of equations in which the polynomial coefficients are derived from the original observations x_i and y_i:

$$\begin{aligned}
\sum_{i=1}^{n} y_i &= \mathrm{a} * n + \mathrm{b} \sum_{i=1}^{n} x_i + \mathrm{c} \sum_{i=1}^{n} x_i^2 \ldots + b_m \sum_{i=1}^{n} x_i^m \\
\sum_{i=1}^{n} x_i y_i &= \mathrm{a} \sum_{i=1}^{n} x_i + \mathrm{b} \sum_{i=1}^{n} x_i^2 + \mathrm{c} \sum_{i=1}^{n} x_i^3 \ldots + b_m \sum_{i=1}^{n} x_i^{m+1} \\
\sum_{i=1}^{n} x_i^2 y_i &= \mathrm{a} \sum_{i=1}^{n} x_i^2 + \mathrm{b} \sum_{i=1}^{n} x_i^3 + \mathrm{c} \sum_{i=1}^{n} x_i^4 \ldots + b_m \sum_{i=1}^{n} x_i^{m+2} \\
&\vdots \\
\sum_{i=1}^{n} x_i^m y_i &= \mathrm{a} \sum_{i=1}^{n} x_i^m + \mathrm{b} \sum_{i=1}^{n} x_i^{m+1} + \mathrm{c} \sum_{i=1}^{n} x_i^{m+2} \ldots + b_m \sum_{i=1}^{n} x_i^{2*\mathrm{m}}
\end{aligned} \qquad (33)$$

More simply written, deleting some obvious indices:

$$\sum Y = a * n + \mathrm{b}\sum X + \mathrm{c}\sum X^2 \ldots + b_m \sum X^m$$
$$\sum XY = \mathrm{a}\sum X + b\sum X^2 + c\sum X^3 \ldots + b_m \sum X^{m+1}$$
$$\sum X^2 Y = \mathrm{a}\sum X^2 + b\sum X^3 + c\sum X_i^4 \ldots + b_m \sum X^{m+2}$$
$$\vdots$$
$$\sum X^m Y = \mathrm{a}\sum X^m + b\sum X^{m+1} + c\sum X^{m+2} \ldots + b_m \sum X^{2*\mathrm{m}}$$

(34)

Solving this set of equations is done by putting the coefficients in a matrix equation:

$$\begin{bmatrix} n & \sum X & \sum X^2 & \sum X^m \\ \sum X & \sum X^2 & \sum X^3 & \sum X^{m+1} \\ \sum X^2 & \sum X^3 & \sum X^4 & \sum X^{m+2} \\ \vdots & \vdots & \vdots & \vdots \\ \sum X^m & \sum X^{m+1} & \sum X^{m+2} & \sum X^{2*\mathrm{m}} \end{bmatrix} . \begin{bmatrix} \mathrm{a} \\ \mathrm{b} \\ \mathrm{c} \\ \vdots \\ b_m \end{bmatrix} = \begin{bmatrix} \sum Y \\ \sum XY \\ \sum X^2 Y \\ \vdots \\ \sum X^m Y \end{bmatrix}$$

(35)

This is solved by matrix inversion (or at least the computer will do that for you!). The result gives us a vector with values for the coefficients of the best fitting equation (*a, b, c, … , b_m*) for which the sum of squares of the residuals ($\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, see Equation 10-12) is minimized.

V. MULTIPLE REGRESSION

V.1 Regression with more than one variable.

Very often situations occur in which more than one variable defines the variation in our observations. For instance, consider a data set on drainage basins of similar size. To estimate flooding risks, you want to analyze the magnitude of peak discharges from these basins. From a basic understanding of hydrology, these peak discharges will depend on several variables, such as:

1. The rainfall intensity and other climatic variables
2. The vegetation cover which may absorb part of the rainfall
3. The permeability of the subsoil which promotes either infiltration of rainfall towards the groundwater or promotes rapid overland flow towards the rivers

In reality, there may be (and will be!) a host of other variables which influence the discharge in this complex system. It is often useful to construct a regression equation which predicts a dependent variable, such as peak discharge, from not just one, but all these variables. This is called **multivariate regression**.

The dependent variable in such a case (e.g. peak discharge magnitude) is again denoted by *Y*, the independent variables (e.g. subsoil permeability, vegetation density etc.) are usually indicated with *X*'s with subscripts: X_1, X_2, X_3, ... X_n. The corresponding multivariate regression equation is than formulated as:

$$Y = a + \mathrm{b} * X_1 + c * X_2 + \ldots + b_n * X_n \qquad (36)$$

In this case, we do not calculate regression equations for each dependent variable separately, but we include all the variables in one equation. As in polynomial regression, we can assess the significance of each X_i term separately to tell which variable contributes to the variation in *Y*.

V.2 Difference between multiple linear regression and polynomial regression

Note that **multiple regression is different from fitting polynomial equations!** In this case, the X values (X_1, X_2, X_3, ... X_n) represent measurements from *different variables*, while in the polynomial equation (Equation 25), all values of *X* represent the same variable. Another important difference is that we are not raising the values of the independent variables to higher powers in this example. Therefore, this special example of the multivariate regression is called **multiple linear regression**. Be very careful to note the difference between this and polynomial regression because the equations (Equation 25 and 35) can be easy to confuse!

V.3 Visualizing multiple linear regression

An equation like Equation 35 does not define a regression 'line', as is the case when only one independent variable is involved. If the equation contains two independent variables, the equation describes a plane in three dimensions, as shown in Figure 16. For more than two independent variables, we cannot visualize or graph the regression equation (Unless you are able to draw in four dimensions…). However, mathematically it is not any problem to have more than two independent variables. If you fill in *n* values for all the X_i in Equation 35, you can compute an estimate for *Y*, irrespective of the number *n*.

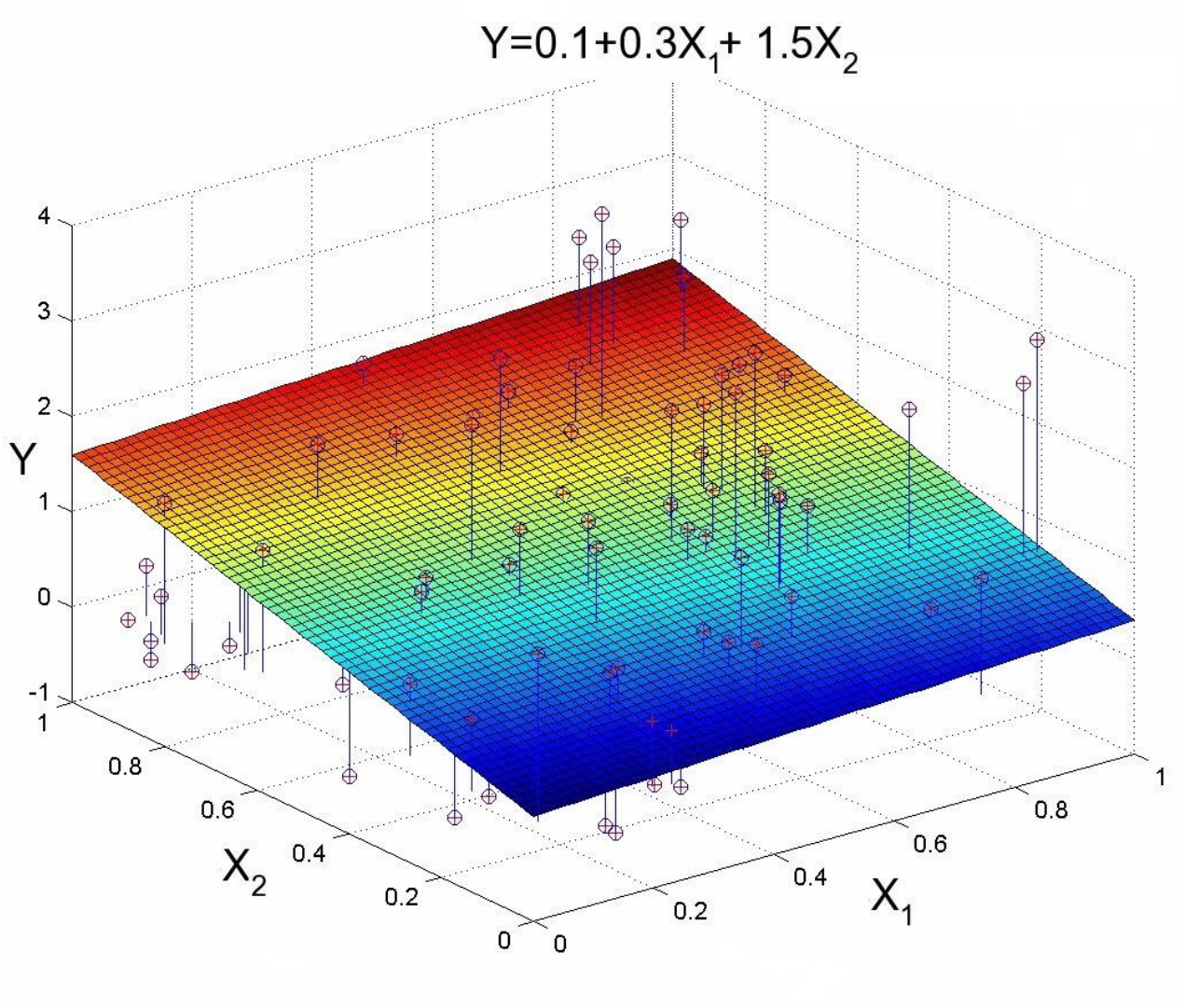

Figure 16: Representation of a regression equation with two independent variables as a plane in three dimensions. The two horizontal axes contain the independent X_1 and X_2 variables, the vertical axis the dependent Y variables. The small circles with red crosses indicate the data points on which the regression was based. The blue vertical lines indicate the deviations of the data points from the regression. Compare with Figure 8 for a regression with only one independent variable.

V.4 Fitting a multiple linear regression

The calculation of the regression equation for a multiple linear regression is quite similar to that of the simple linear regression and that of the polynomial regression. The coefficients are derived from the original observations x_i and y_i, in the same way, the formula just gets longer. Below, we will walk through an example to show you how this type of regression is done. You do not need to be able to reproduce this calculation, because a computer software like Python will do it for you. However, seeing the steps will help you to better understand how this regression works, especially now that the mathematics gets more abstract because we are adding multiple variables. I will therefore place the full calculation in Section V.7 Extra reading: The mathematics behind a multiple linear regression

With multivariate regression calculation using this method, we cannot just put the independent *X* variables in any random order. The *X* variables must be in the order of strongest correlation with *Y*. To determine the order, we must determine the correlation coefficients (Pearson's r) of every independent *X* variable with *Y* and take their absolute values. The variable with the highest absolute

correlation becomes X_1, the next highest X_2, etc. For instance, we have determined the highest discharge peak of several river basins of similar size as dependent variable *Y*, and the percent forest cover, the average subsoil permeability, and the rainfall intensity as *X* variables. We start with calculating the correlations between *Y* and all the X's. The correlation coefficients are given in Table 6.

Table 6: Correlation coefficients between independent variables and dependent variable in a dataset about river discharge.

	% forest cover	***subsoil permeability***	***rainfall intensity***
discharge peak	0.4	-0.6	0.7

The order of strongest correlation is *rainfall intensity* > *subsoil permeability* > *forest cover.* So rainfall intensity should become X_1, subsoil permeability becomes X_2, and forest cover becomes X_3.

V.5 Significance of a multiple linear regression

Again, we can test the significance of the complete regression equation and the contribution of each variable separately. If we want to do this, we must also compute the regression with only X_1, the regression with X_1 and X_2 as independent variables, the regression with X_1, X_2, X_3, etc. By adding variables to the equation step by step and computing the relevant sums of squares SS_R of the regression, we can find out which variables contribute significantly to the regression and decide how many variables we should include. For the regression with only one variable, this results in a SS_{R1}, for that with two variables in SS_{R2}, three variables SS_{R3} etc. Table 7 shows the complete ANOVA table worked out for the first three variables:

Table 7: ANOVA table for a multiple linear regression with m independent variables and n observations. The fourth and fifth lines highlight the calculations of the significance of adding the second and third most important independent variables (X_2 and X_3).

Source of variation	***Sum of squares***	***Degrees of freedom (df)***	***Mean squares***	***F-test***
Regression all *m* variables	SS_R	*m*	$MS_R=SS_R/df$	MS_R/MS_D
Regression with only X_1	SS_{R1}	*1*	$MS_{R1}=SS_{R1}/df$	MS_{R1}/MS_{E1}
Addition due to X_2	$SS_{R1-2}=SS_{R2}-SS_{R1}$	*1*	MS_D	MS_{R1-2}/MS_D
Addition due to X_3	$SS_{R2-3}=SS_{R3}-SS_{R2}$	*1*	MS_D	MS_{R2-3}/MS_D
.....	.....	.....	.....	.....
Deviation from regression	SS_E	*n - m - 1*	$MS_E=SS_E/df$	
Total variation	SS_T	*n-1*		

V.6 Complications with multiple linear regression and significance testing

The procedure of adding or removing independent variables one by one is sometimes referred to as

stepwise regression (or **stepwise linear regression** in this case). The most logical way to use it is to start with the most strongly correlating independent variable (X_1) and work your way up (adding X_2, X_3, X_4 etc). However, the method is not foolproof, because it is possible in theory that one will miss combinations of independent variables which explain a large part of the variability in the dependent variable. For example, the combination of X_1, X_3 and X_4 may be a better predictor of Y than the combination of X_1 and X_2. However, by strictly following the stepwise procedure in order (X_1, X_2, X_3, ... X_n), you might miss this combination. The solution would be to test every possible combination of independent variables. As you can imagine, this requires an exponentially increasing number of calculations when datasets contain more and more independent variables.

A problem in calculating the regression equations in polynomial and multivariate regression can be the large numbers that may result from the sums of squares. For example, if you have 50 observations expressed in large numbers in the order of hundreds or thousands, the sums of squares of quadratic and higher order terms will be enormous. This will lead to rounding errors in the computer. Therefore, computer programs for regression calculations usually subtract the mean of each variable from the individual observations. The observations are thus expressed as deviations from the mean.

The stepwise regression method used to be a popular way to find meaningful relationships in large datasets, but it has come under scrutiny because of the way it uses significance testing. If we allow ourselves to test increasingly large combinations of independent variables while using a 95% confidence limit ($p = 0.05$) to judge whether a combination of independent variables significantly explains our dependent variable, we greatly increase our chance of finding a significant result by accident. We call such an accidental result a **false positive**. The procedure of testing a large number of combinations of variables or explanations to find a "significant" result is sometimes referred to as **p-hacking**. This article[1] makes a strong case against the uninformed use of stepwise regression.

Overfitting is also a risk in multiple linear regression. The more independent variables we add to our dataset, the higher the chance that we obtain a good fit to our test data (a high R^2). This is why it is important to be conservative when adding variables to our datasets. Remember always that data analysis is a tool, and that it does not do the interpretation for you. Always consider whether you can reasonably assume that an independent variable might be *causally* related to the dependent variable before you add it to your dataset. The *spurious correlation* examples in II. CORRELATION show why it is important to not blindly trust statistics to help you find meaningful relationships in (large) datasets!

[1] Peter L Flom et al., "Stopping Stepwise: Why Stepwise and Similar Selection Methods Are Bad, and What You Should Use," *NESUG 2007*, 2007, 1–7.

V.7 Take Home Messages

- When we want to estimate one dependent variable with multiple independent variables, we need to do a **multiple regression**.
- In this syllabus, we only deal with **multiple linear regression**, but note that you can also combine multiple independent variables using non-linear relationships.
- Fitting a multiple regression can be done with **ordinary least squares,** similar to simple linear regression, and we can use an ANOVA to find out if variables contribute significantly to the prediction.
- **Stepwise (linear) regression** is a method for finding combination of independent variables that predicts a dependent variable, but it is highly sensitive to **p-hacking** and **overfitting** and should be applied with caution.
- Always consider whether an independent variable has a logical, causal relationship with the dependent variable before adding it to a multiple regression to avoid **spurious correlations**.

V.7 Extra reading: The mathematics behind a multiple linear regression

Let's look at the calculations underlying the fit of a multiple linear regression. Suppose we have *m* independent variables, and *n* observations of the dependent and independent variables. The independent variables are stored in a matrix of *n* rows and *m* columns. The dependent variable is stored in its own matrix, which has *n* rows (one for each observation) and one column. The calculation for finding the optimal set of parameters of the multiple linear regression starts with calculating **cross products** between the variables. In other words: We are multiplying each observation *x* with each corresponding *y*, and each *x* with one of the other *x*-es. The mathematical representation of such a matrix cross product is given below:

$$\begin{bmatrix} x_{1,1} & x_{2,1} & x_{3,1} & \dots & x_{m,1} \\ x_{1,2} & x_{2,2} & x_{3,2} & \dots & x_{m,2} \\ x_{1,3} & x_{2,3} & x_{3,3} & \dots & x_{m,3} \\ \vdots & \vdots & \vdots & & \vdots \\ x_{1,n} & x_{2,n} & x_{3,n} & \dots & x_{m,n} \end{bmatrix} \times \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}$$

(37)

Once again, we are not fully deriving the origin of these matrix multiplication steps here, but feel free to check out the details here if you are interested! The sums of the cross-products of these two matrices for one observation (one row in the matrix) looks something like this:

$$x_{i,1} * y_i + x_{i,2} * y_i + \cdots + x_{i,j} * y_i + x_{i,1} * x_{i,1} + x_{i,1} * x_{i,2} + \cdots + x_{i,j} * x_{i,m} + x_{i,m} * x_{i,m}$$

(38)

Here, *i* denotes a row of the matrix of all observations of the independent X variables, and *j* a column. Similar to the linear and polynomial regressions, this results in a set of equations with the coefficients of the regression line as unknowns:

$$\sum_{i=1}^{n} y_i = \mathrm{a} * n + \mathrm{b}\sum_{i=1}^{n} x_{i,1} + \mathrm{c}\sum_{i=1}^{n} x_{i,2} + \cdots + b_m \sum_{i=1}^{n} x_{i,m}$$
$$\sum_{i=1}^{n} x_{i,1} y_i = \mathrm{a}\sum_{i=1}^{n} x_{i,1} + \mathrm{b}\sum_{i=1}^{n} x_{i,1}^2 + \mathrm{c}\sum_{i=1}^{n} x_{i,1}x_{i,3} + \cdots + b_m \sum_{i=1}^{n} x_{i,1}x_{i,m}$$
$$\sum_{i=1}^{n} x_{i,2} y_i = \mathrm{a}\sum_{i=1}^{n} x_{i,2} + \mathrm{b}\sum_{i=1}^{n} x_{i,2}x_{i,1} + \mathrm{c}\sum_{i=1}^{n} x_{i,2}^2 + \cdots + b_m \sum_{i=1}^{n} x_{i,2}x_{i,m}$$
$$\vdots$$
$$\sum_{i=1}^{n} x_{i,m} y_i = \mathrm{a}\sum_{i=1}^{n} x_{i,m} + \mathrm{b}\sum_{i=1}^{n} x_{i,1}x_{i,m} + \mathrm{c}\sum_{i=1}^{n} x_{i,2}x_{i,m} + \cdots + b_m \sum_{i=1}^{n} x_{i,m}^2$$

(39)

or, more simply written, deleting some obvious indices:

$$\sum Y = a * n + \mathrm{b}\sum X_1 + \mathrm{c}\sum X_2 + \ldots + b_m \sum X_m$$
$$\sum X_1 Y = \mathrm{a}\sum X_1 + \mathrm{b}\sum X_1^2 + \mathrm{c}\sum X_1 X_2 + \ldots + b_m \sum X_1 X_m$$
$$\sum X_2 Y = \mathrm{a}\sum X_2 + \mathrm{b}\sum X_2 X_1 + \mathrm{c}\sum X_2^2 + \ldots + b_m \sum X_2 X_m$$
$$\vdots$$
$$\sum X_m Y = \mathrm{a}\sum X_m + \mathrm{b}\sum X_m X_1 + \mathrm{c}\sum X_m X_2 + \ldots + b_m \sum X_m^2$$

(40)

Solving this set of equations is done by putting the coefficients in a matrix equation:

$$\begin{bmatrix} n & \sum X_1 & \sum X_2 & \cdots & \sum X_m \\ \sum X_1 & \sum X_1^2 & \sum X_1 X_2 & \cdots & \sum X_1 X_m \\ \sum X_2 & \sum X_2 X_1 & \sum X_2^2 & \cdots & \sum X_2 X_m \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \sum X_m & \sum X_m X_1 & \sum X_m X_2 & \cdots & \sum X_m^2 \end{bmatrix} \cdot \begin{bmatrix} a \\ \mathrm{b} \\ \mathrm{c} \\ \vdots \\ b_m \end{bmatrix} = \begin{bmatrix} \sum Y \\ \sum X_1 Y \\ \sum X_2 Y \\ \vdots \\ \sum X_m Y \end{bmatrix}$$

(41)

This matrix is solved by matrix inversion using a computer software.

VI. ALTERNATIVE SOLUTIONS TO THE REGRESSION PROBLEM: *Reduced major axis (RMA) and principal axis.*

VI.1 Problems with Ordinary Least Squares regression

We now return to simple linear regression with one independent variable, to illustrate some alternative approaches that are often useful. Two problems may arise with the OLS regression method discussed in Section III.2 How to find the right line?:

Firstly, in Section III.2 How to find the right line?, we assumed that the random errors in the data are only in the observations of the dependent variable *Y*, and that the independent variable *X* is known infinitely accurately. Remember that, in the regression model stated in Equation 14 ($y_i = a + b * x_i + \varepsilon_i$), the error term is only attributed to y_i. Very often, this is not the case, and errors may be present also in the observations of x_i. Consider for instance a curve of past sea level rise that is reconstructed from ancient sea level indicators, such as the tops of coral reefs, and radiometric datings (based on carbon-14 or uranium series dating) of the carbonate in the last-grown parts of these corals (see this paper[1] for examples if you are interested in this type of research). In this case, both the datings and the sea level indicators are subject to errors. Every radiometric dating has an uncertainty and the relation of the height of the coral reef to former sea level is not very exact. In that case, you would like to use a regression method that accounts for errors in both the dependent and independent variable.

Secondly, it is often not very certain which variable should be the independent one and which the dependent one. Very often, this is known from *a priori* knowledge of the processes. For example, the amount of sediment transported over a riverbed is always dependent on the river discharge, not the other way round. The causal order of the variables is also sometimes known from the definition of the problem. For example, if you want to analyze real estate prices as a function of time, time is by definition the independent variable (time is almost always an independent variable, especially in geosciences, but more on that later). However, in many cases it is not so clear which variable is the cause and which variable responds. For instance, in certain types of clastic sedimentary sequences there is a relation between grain size of the sediment and the sedimentation rate (the thickness of sediment deposited per unit time). If you want to quantify this relation, it is nonsense to say that courser sediment *causes* a higher sedimentation rate, and therefore should be the independent variable. However, the same holds for sedimentation rate: It does not cause the grains to be larger. In this case there is no theoretical reason why one of the variables should be the independent and which the dependent one. The problem is that using the OLS regression method in Section III.2 How to find the right line? we obtain different regression lines depending on which variable we choose as the independent variable. See **Error! Reference source not found.** for an example.

[1] Fiona D. Hibbert et al., "Coral Indicators of Past Sea-Level Change: A Global Repository of U-Series Dated Benchmarks," *Quaternary Science Reviews* 145 (August 1, 2016): 1–56, https://doi.org/10.1016/j.quascirev.2016.04.019.

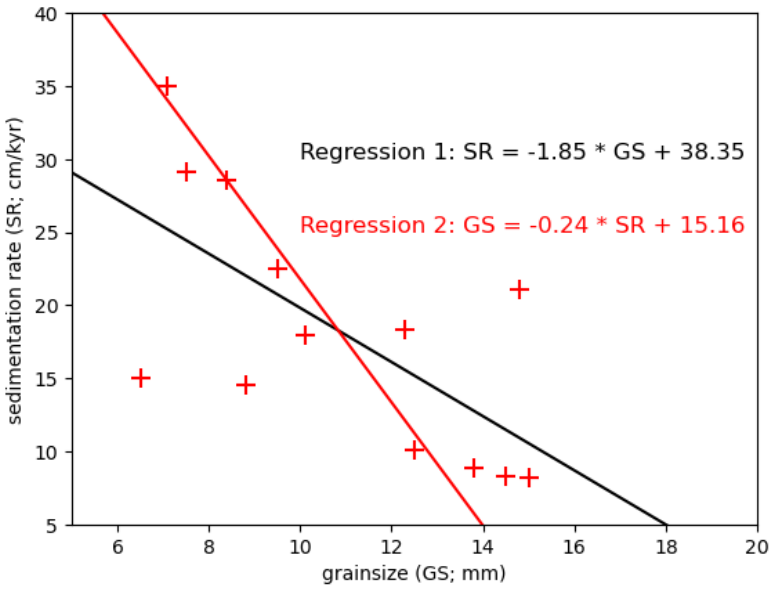

Figure 17: Two different regression lines on the same bivariate data set depending on the choice of the independent variable. For Regression 1, grain size is taken as the independent variable. For Regression 2, sedimentation rate taken as the independent variable.

VI.2: Alternative ways to minimize the distance between datapoint and regression

There are alternative methods that overcome these problems. First, consider the way the errors can be treated in the calculation of the regression line. As we know, the coefficients of the regression equation are found by minimizing the deviations of the data points from the regression line. This can be done in several ways (see Figure 18).

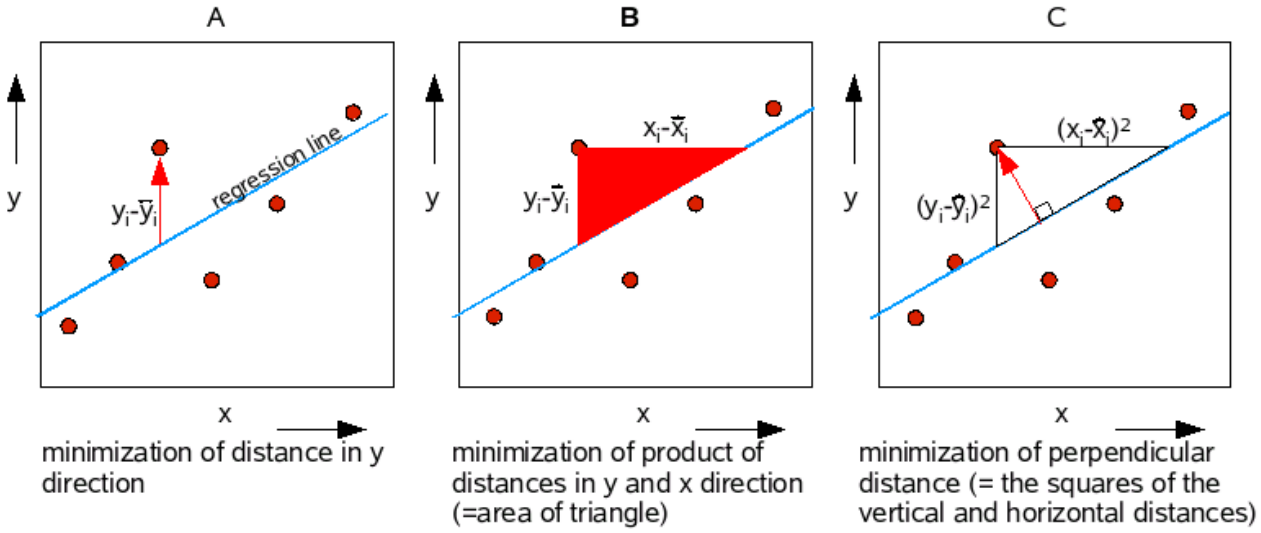

Figure 18: Different criteria for minimizing the deviations of the data points from a regression line. The method of A is applied in the Ordinary Least Squares (OLS) simple linear regression method presented in Section III.2 How to find the right line?, B is applied in the reduced major axis (RMA) regression method, C in the principal axis (PA) regression method (see text).

The simplest way to solve this issue is by minimizing only the difference of the dependent variable *y* with the regression line, assuming that the deviations (*D*) are entirely due to errors in the observations of *y:*

$$D = y_i - \hat{y}_i$$

(42)

This is what we did in Section III.2 How to find the right line? when we applied the Ordinary Least Squares regression method. The criteria in B and C of Figure 18 also account for errors in the independent variable *x*. In B this is done by multiplying the deviations in the *x* and *y* direction with each other:

$$D = (y_i - \hat{y}_i) * (x_i - \hat{x}_i)$$

(43)

This results in a deviation that is proportional to the area of a triangle, formed by the regression line and the deviations in the *x* and *y* directions. We can also minimize the deviation in a direction perpendicular to the regression, as is done in C. In that case, this amounts to calculating the summed squares of the deviations in *x* and *y* direction:

$$D = (y_i - \hat{y}_i)^2 + (x_i - \hat{x}_i)^2$$

(44)

VI.3: Reduced Major Axis Regression (RMA)

The method applied in B of Figure 18 results in a regression line known as the *reduced major axis (RMA)*. The calculation of the coefficients of that line is simple. First, you need to calculate the standard deviations, covariance and means of the variables. The slope of the regression line is then:

$$b = \frac{\sigma_y}{\sigma_X}$$

(45)

The sign of the slope *b* (plus or minus) is the same as that of the covariance (the Pearson's r value, see Equation 1). The intercept is found by:

$$a = \bar{Y} - b * \bar{X}$$

(46)

In Davis (2002[1]) significance tests are described for the slope and intercept of the RMA regression line. We will not derive these here, because the details of RMA are not part of the learning goals for this course. For now, it is sufficient to know that this alternative method exists.

VI.4: Principle Axis Regression

The method applied in C of Figure 18 is more intricate. The resulting regression line is known as the *principal axis* or *major axis.* Calculation of the slope coefficient requires matrix algebra and is not discussed here. However, it is interesting to see (and important for the next chapter) what these regression lines mean graphically. This is demonstrated in Figure 19.

[1] Davis and Sampson, *Statistics and Data Analysis in Geology*.

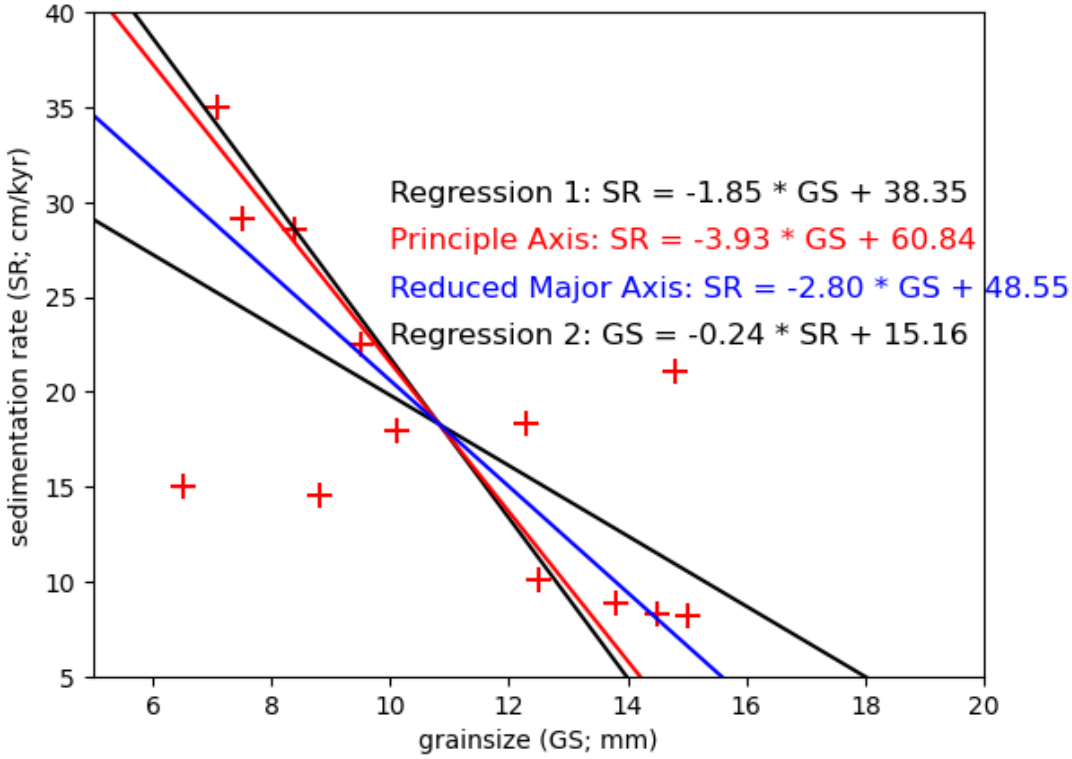

Figure 19: Regression lines of sedimentation rate on grainsize and grainsize on sedimentation rate (black lines, compare Figure 18) and reduced major axis (blue) and principal axis (red) based on the same data.

In Figure 19, again both the regression lines of sedimentation rate on grainsize and grainsize on sedimentation rate (black lines) of the same data as in Figure 18. The cosine of the angle between the lines is numerically equivalent to correlation coefficient between *X* and *Y*. The reduced major axis line (RMA; in blue) and the principal axis line (red) bisect exactly the angle between the two blue regression lines. All lines cross each other at the means of *X* and *Y*. The principal axis line is oriented in such a way that the variance of the data is maximal along this axis. If we would rotate Figure 19 around the mean of *X* and *Y* until the principal axis lies horizontal, the spread between the data points in the horizontal direction proves to be the largest possible. In the vertical direction the variation is smallest (Figure 20). We will make extensive use of that property of the principal axis later when we discuss Principle Component Analysis (or Factor Analysis).

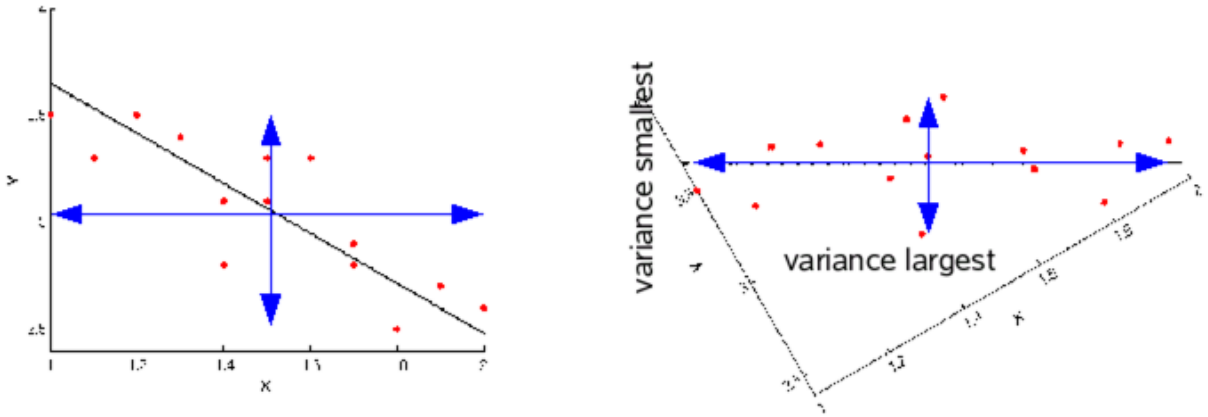

Figure 20: Example of a dataset unrotated (left) and rotated (right) around the mean of X and Y. Left, the principal axis is rotated to 0 degrees to show that the spread of the data, and the variance is largest along the principal axis (horizonal), and smallest perpendicular to the principal axis (vertical). The variance along the principal axis is larger

than along the original x and y axes (compare length of arrows left and right).

VI.5 Take Home Messages

- Ordinary Least Squares regression assumes that all the sources of error in our regression analysis originate from uncertainty in data on the independent variable(s).
- Ordinary Least Squares regression yields a different regression result based on which variable is considered independent. It is therefore important to have a good reason for this choice.
- To consider uncertainty in the dependent variable while performing a regression, we can apply Reduced Major Axis regression (RMA) or Principle Axis regression.
- Principle Axis regression maximizes the variability in the dataset in the direction along the regression line. It forms the basis for Principle Component Analysis.

VII NON-CONTINUOUS VARIABLES AND LOGISTIC REGRESSION

VII.1 Continuous, discrete, and binary variables

In the preceding sections we have made acquaintance with various types of regression problems based on datasets with single or multiple independent variables and relationships with different shapes. What all these problems have in common is that the variables we were working with are **continuous variables**. A continuous variable can meaningfully take any value in a certain range. For example, time is a continuous variable: We can cut the range of the variable up into smaller or larger intervals (from milliseconds to millions of years) and it will still make sense. There is no point on the unfathomably long timeline of the universe that is somehow, by itself, more "logical" than any other moment in time. In addition, we can add and subtract increments of time from each other in a linear way to obtain new moments in time which are equally plausible. In other words: There are no "gaps" in the time variable we must avoid. Other examples of continuous variables include distance, mass, length, voltage, or velocity. Continuous variables are **uncountable**, it is impossible to name all values they can attain, even if you would have infinite time. Continuous variables are often the outcomes of some sort of measurement; They are **measurable**. We encounter then a lot in geosciences!

Not all variables are like this. Some variables can only take on certain values. An example is the outcome of a survey or an election. In the Netherlands, there is a limited number of political parties (which has been growing lately) you can vote for. When you cast your vote, it makes sense to vote for Party 1 or Party 2, but it does not make sense to vote for Party 1.7 or Party π. The list of parties in the election is a **discrete variable**, also called a **categorical variable**. Other examples of discrete variables include color, gender, nationality, species, or geological period. All these variables can only take a limited number of values. Discrete variables are **countable**. In theory, all the values a discrete variable can take can be counted, although in some cases the number can be large. Everyone who has been to a hardware store to buy paint knows how many names we have for different colors, but the list is not endless, so in theory you should be able to write all of them down if you had enough time (and were really bored!). Contrary to continuous variables, it is not generally possible to measure the value of a discrete variable. This is a bit confusing because we do have tools to measure color (called colorimeters). However, what these tools actually measure is the wavelength of light that gets reflected of the surface of the object we want to determine the color of. It does not measure the "category" of the color (e.g. "red" or "violet"). This light wavelength can in fact take on an infinite number of values and is therefore a continuous variable. The color category is **unmeasurable** and is a discrete variable.

There is a special group of discrete variables which can only take up two values. We call these **binary variables**. Examples of binary variables include "yes or no", "win or lose", "dead or alive", and "0 or 1" (in the context of, for example, computer bits). Binary variables (and discrete variables in general) are in fact surprisingly rare in nature, which loves grey areas and transitions. However, they occur frequently in datasets and require special treatment when included in statistical models such as regressions.

VII.2 Dummy variables

The first and most straightforward instance in which we can encounter a discrete or binary variable is as the independent variable in our dataset. An example of this is provided in Figure 21. This figure displays a dataset which results from an experiment in which the height of people on the street in Amsterdam was measured and the same people were asked whether they were born in the Netherlands or came from another country.

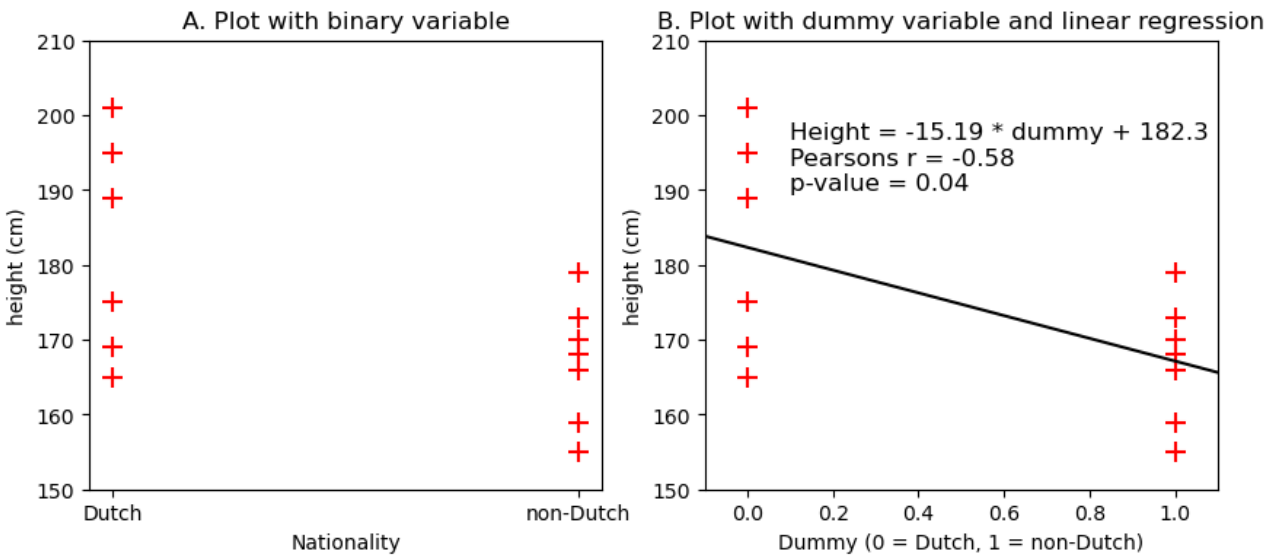

Figure 21: Example of a dataset with a binary independent variable (Nationality) and a continuous dependent variable (height in cm; plot A). In Plot B, the independent variable is replaced by a dummy variable to allow us to carry out a simple linear regression and test whether the binary variable significantly predicts the dependent variable (it does).

If we want to know whether the height the people is significantly correlated to their nationality (Dutch or non-Dutch), we have two options:

Firstly, we can apply the Student's t-test or Welch's t-test (depending on whether we believe the variances of the two groups are equal). You have learned how to do this in the Statistics part of the course, so we will not go in to details here.

Secondly, we can apply a linear regression to find out if the difference between the groups is statistically significant. Since we cannot do a linear regression on a discrete variable, we need to transform our independent variable into a **dummy variable**. To do this, we simply replace the categories of the independent variable with numbers. By default, we choose the numbers 0 and 1 for this. Choosing these numbers makes our regression result easier to interpret, as will be explained below. You can see how this works for our dataset in Panel B of Figure 21.

After creating the dummy variable, we can calculate the Pearson's r value of the correlation, which tells us the direction of the effect of our independent variable (nationality) on our dependent variable (height). In this case, a negative Pearson's r means that the Dutch people who were measured are on average taller than the non-Dutch people in our dataset. More importantly, we can do an F-test and calculate a p-value for our simple linear regression. The result of this shows that the Dutch people are indeed significantly taller than the non-Dutch people within a 95% confidence level ($p < 0.05$). Finally, the parameters of our regression result (slope and intercept) now tell us what the size of the effect is. Since we chose 0 and 1 as our dummy variable, the slope (-15.19) shows us that the Dutch people were on average 15.19 cm taller than the non-Dutch people, and the intercept (182.3) tells us that the average height of the Dutch people was 182.3 cm. We can then easily calculate the average height of the non-Dutch group by adding up the slope and intercept (-15.19 + 182.3 = 167.11 cm).

Note that this regression is only meaningful for x-values of 0 or 1. After all, a value between or outside these categories does not really make sense (you cannot be born half within the Netherlands and half outside the Netherlands).

VII.3 Discrete dependent variables

The example in Section VII.2 Dummy variables is rather straightforward and might make you think that dummy variables are the solution to all our discrete variables. Unfortunately, this is not the case. Sometimes, the dependent variable is a discrete variable, and in this case replacing it with a dummy variable will not help us. To explain why, consider the example in Figure 22.

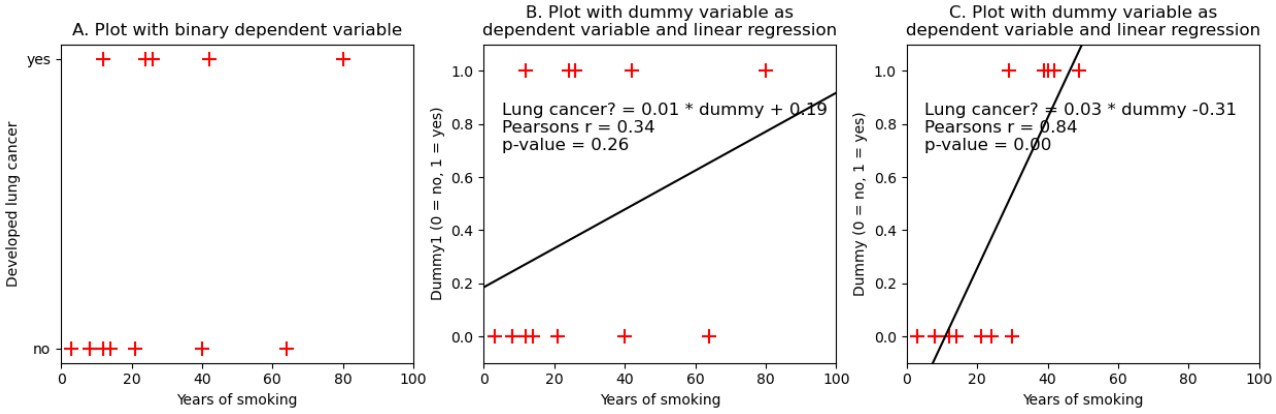

Figure 22: Example of a dataset with a continuous independent variable (years of smoking) and a binary dependent variable (developed lung cancer; plot A). In Plot B and C, the dependent variable is replaced by a dummy variable in an attempt to carry out a simple linear regression and test whether the binary variable significantly predicts the dependent variable. The result is not meaningful.

In Figure 22, we plot data of patients' smoking history (some of them started very early!) against their medical outcome; in this case a test whether the patient developed lung cancer (plot A). In this case, the dependent variable (whether the patient developed lung cancer) is binary. We know there is on average a higher risk of developing lung cancer the longer a patient smokes, so our first intuition might be to use a linear regression to test whether this is the case using a dummy variable for the medical outcome (plot B). Statistically, this can be done, but the result looks very suspicious: All the non-cancer patients' datapoints lie below the line, and all the of data of patients with cancer lies above it. This seems like a very biased result, and it suggests that the line underestimates the risk of long cancer for patients that did develop it and overestimates it for non-cancer patients. In addition, consider how we can interpret this result: The line never reaches the point of zero or one, while all the data that goes into the regression takes one of these values. This makes it difficult to say what any outcome of the dependent variable means. After all, it is not possible to have 40% lung cancer… Plot C looks better from a statistical point of view (it also has a lower p-value and higher Pearson's r), but the regression is even harder to interpret, because the fitted line crosses the horizontal lines at y = 0 and y = 1. This is completely meaningless, because one cannot have "negative lung cancer" or "more than 1" lung cancer. In conclusion: It seems that a linear regression is not a fitting solution for this type of data.

VII.4 Logistic regression

To find a better fit for this type of data, we need to find a curve that "hugs" the bottom ("no lung cancer") and top ("lung cancer") of the plot without crossing the lines at y = 0 and y = 1. The type of function that does this is called a **sigmoid function**, also called the **logistic function**, and it is described by the following formula:

$$y(x) = \frac{1}{1 + e^{-x}} \tag{47}$$

A plot of the standard logistic curve is given in Figure 23.

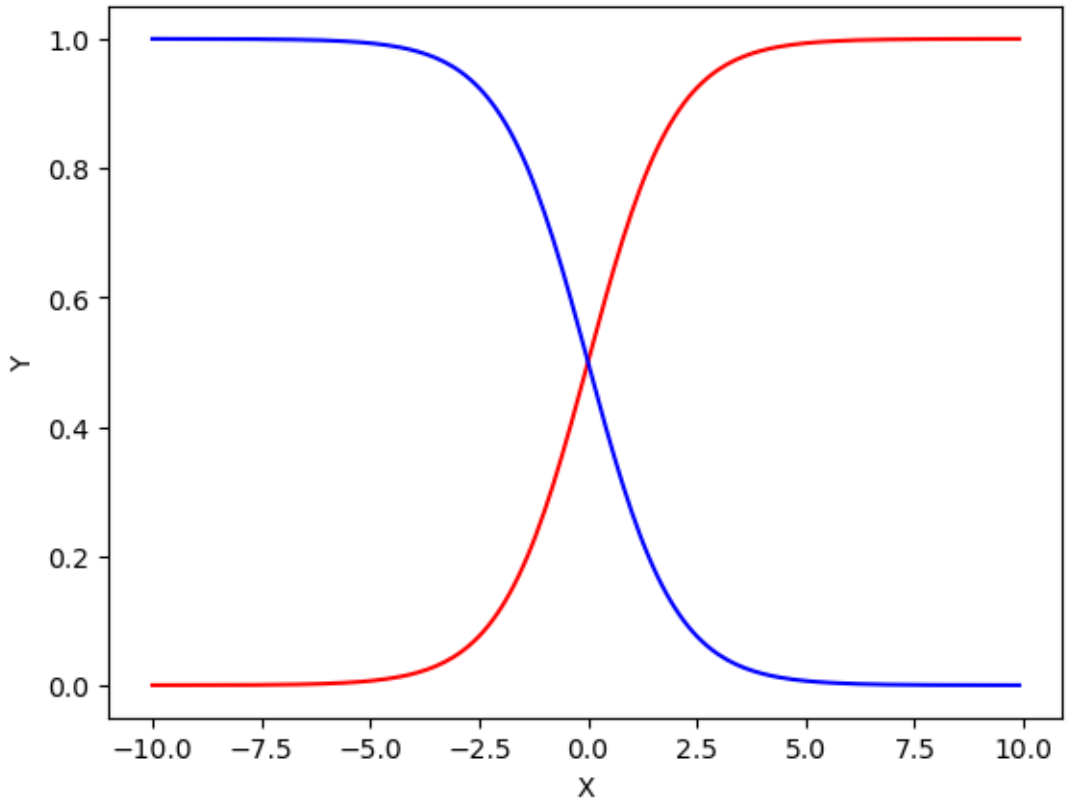

Figure 23: The standard logistic curve (or sigmoid function; in red) and its negative counterpart (in blue)

As you will see, this curve has exactly the shape we need for the properties of our data with a binary dependent variable: It hugs the lines y = 0 and y = 1, but never crosses it for any x-value and it is also shaped in such a way that it stays very close to the datapoints with a low x value (and y = 0) and the datapoints with a high x value (for which y is 1). By the way, if the effect would be opposite (high x values yield y = 0 and vice versa), then we can simply invert the sigmoid curve by taking 1 minus the positive sigmoid, as can also be seen in Figure 23.

The properties of the logistic curve are no coincidence because it is directly derived from a distribution with a very similar shape to the **normal distribution** (which you know well by now) and describes the **cumulative probability** of an event (y) given the size of the independent variable (x) that causes the event. The logistic function belongs to the family of **cumulative density functions**, and its derivative is the normal distribution. Another property of the logistic function is that it is the so-called "maximum entropy" solution to the problem of converting a real number to a probability. What this means is that it is the solution to our problem of dealing with a binary dependent variable with the fewest assumptions about the relationship between the variables. The last few sentences may seem confusing, and during this course we will not go into the details of what this means exactly and why this solution is best. In the VII.7 Extra reading: The similarities between the logistic curve and the normal distribution section, we will show how we can take the derivative of the logistic function to show that the underlying data distribution is very similar to the normal distribution. This blog post also clearly explains the relationship between these two functions. Tor now, the only thing you need to keep in mind is that this function can be used in a regression that takes a continuous independent variable (x) and predicts the probability of a binary outcome (y). We call this type of regression a **logistic regression**.

Logistic regression is incredibly common in data science because the outcome of a dataset with predictors (independent variables) is very often a binary probability. A medical test, such as our lung cancer example above, is a common example, but there are many more, such as:

- Estimating the probability that a voter votes for one of two parties in the US elections based on their age, sex, State, profession, previous voting behavior, voting behavior of their parents, ethnicity, etc.
- The chance that a driver will need to use their car insurance based on their age, sex, driving history, number of kilometers driven per year, type of car, city of residence, etc.
- The chance that an extreme weather event (storm, flood, heatwave, etc) will strike depending on the location's latitude and longitude, mean annual temperature, weather history, elevation, distance to the sea, etc.

As you can see from the examples above, many of these applications have very high stakes and therefore involve a lot of money!

VII.5 Fitting a logistic function

In this course, we will not manually fit a logistic function. We have Python to do this for us! However, it is useful to know how this fitting works approximately. The general model we want to fit in a logistic regression is the following:

$$\log\frac{p}{1-p} = a + \mathrm{b} * X_1 + c * X_2 + \ldots + b_n * X_n \qquad (48)$$

In this formula, the term $\log\frac{p}{1-p}$ is called the **log-odds**. It is the natural logarithm of the *probability* that is the dependent variable (*p*; divided by its opposite: 1 - *p*). In the end, we want to know this value *p* for each set of values for the independent variables *x* ($X_1, X_2, X_3, \ldots X_m$). In other words, *p* is our dependent variable, and it can take any value between 0 and 1. The part on the right side of the equation is identical to the equation for the V. MULTIPLE REGRESSION (Equation 36). Just like in multiple linear regression, the values for $X_1, X_2, X_3, \ldots X_m$ represent the values for each of the independent parameters we use to estimate *p*. The parameters ($a, b, c, \ldots, b_n$) are the parameters we want to know to be able to estimate *p* for each set of values $X_1, X_2, X_3, \ldots X_m$.

Firstly, it is important to note that, unlike OLS regression, there is no *closed form* expression for the optimal fit of a logistic function. This means that we cannot write a formula in which we can feed the data for $X_1, X_2, X_3, \ldots X_m$ and values for *p* and which gives us one value for the optimal parameters of the logistic function. Instead, the logistic function is often fitted using a process called **Maximum Likelihood Estimation**, in which the software makes a "first guess" of the parameters and then progressively updates its guesses to get the optimum values for the parameters. This updating is mathematically complex and will not be explained in detail here. If you want to read more about how this works, this page explains the steps and calculations.

While fitting a logistic regression, we minimize the **log-likelihood** (*l*) of the probability (*p*) we are trying to estimate. The function of this log-likelihood is given below:

$$l_i = \begin{cases} -\log(p_i), & if\ y_i = 1 \\ -\log(1 - p_i), & if\ y_i = 0 \end{cases} \qquad (49)$$

When you fit a logistic regression using Python, the software is trying to make this value as small as possible (0 is perfect, 1 is infinitely bad) by changing the parameters ($a, b, c, \ldots, b_n$) iteratively. You

can see the result of this process applied on our example data from Figure 22 in the figure below.

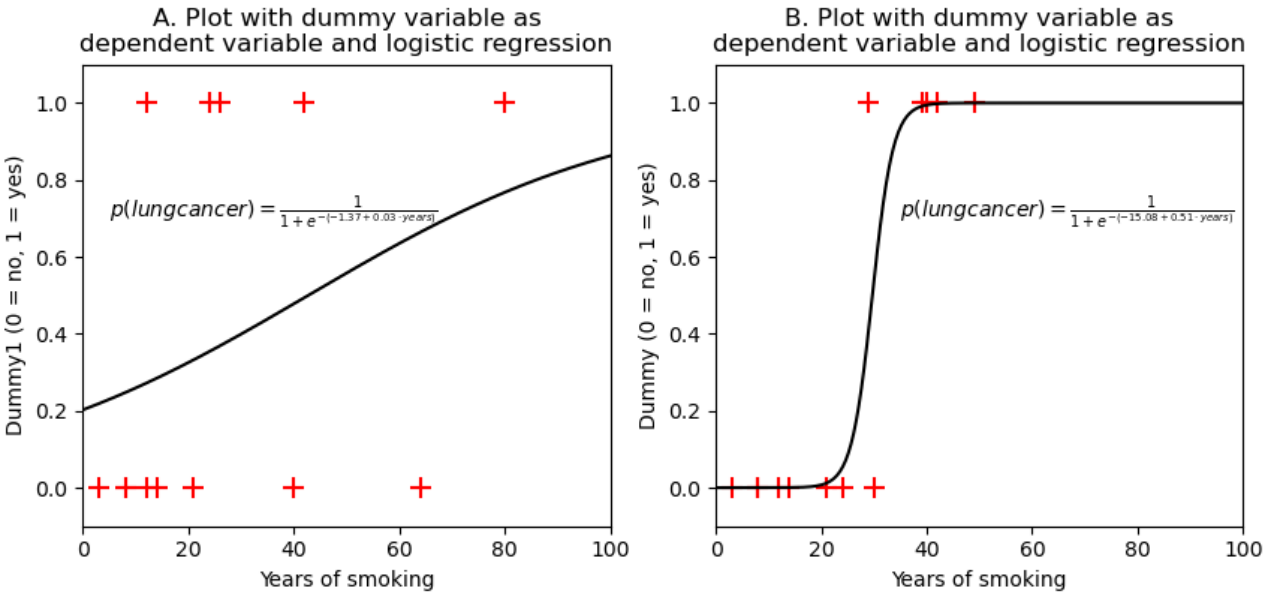

Figure 24: Example of the two dataset from Figure 22 with a continuous independent variable (years of smoking) and a binary dependent variable (developed lung cancer). The dependent variable is replaced by a dummy variable to carry out a logistic regression. The formulas describing the resulting logistic curves are given.

VII.6 Take Home Messages

- Some variables in our regression are discrete, meaning they can only take certain values
- When a variable can only take two values, we call it binary
- Discrete independent variables can be replaced by a dummy variable
- For discrete dependent variables, regular regression methods do not work. In those cases we need to apply a logistic regression to estimate the probability of the dependent variables as a function of the independent variables in a meaningful way.

VII.7 Extra reading: The similarities between the logistic curve and the normal distribution

To understand how the logistic function is related to the normal distribution, it helps to see how the derivative of the logistic function (a cumulative density function) one can be used to derive the formula of the underlying probability density function. We start with the logistic function:

$$y(x) = \frac{1}{1 + e^{-x}} \tag{50}$$

Now we take it's derivative in a few steps:

$$\frac{dy}{dx}(x) = \frac{d}{dx}\left(\frac{1}{1 + e^{-x}}\right) = \frac{(1 + e^{-x}) \cdot 0 - 1 \cdot (-e^{-x})}{(1 + e^{-x})^2} = \frac{e^{-x}}{(1 + e^{-x})^2} \tag{51}$$

This can be rewritten as:

$$\frac{dy}{dx}(x) = \frac{e^{-x}}{(1+e^{-x})^2}$$
$$= \frac{e^{-x}}{(1+e^{-x})(1+e^{-x})}$$
$$= \frac{e^{-x} * e^{x}}{(1+e^{-x})(1+e^{-x}) * e^{x}}$$
$$= \frac{1}{(1+e^{x})^2}$$

(52)

The normal distribution is written as:

$$y(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$

(53)

The thing these two distributions have in common is that they are both a function of the term e^{-x^2}, and therefore have a similar shape. The figure below illustrates this.

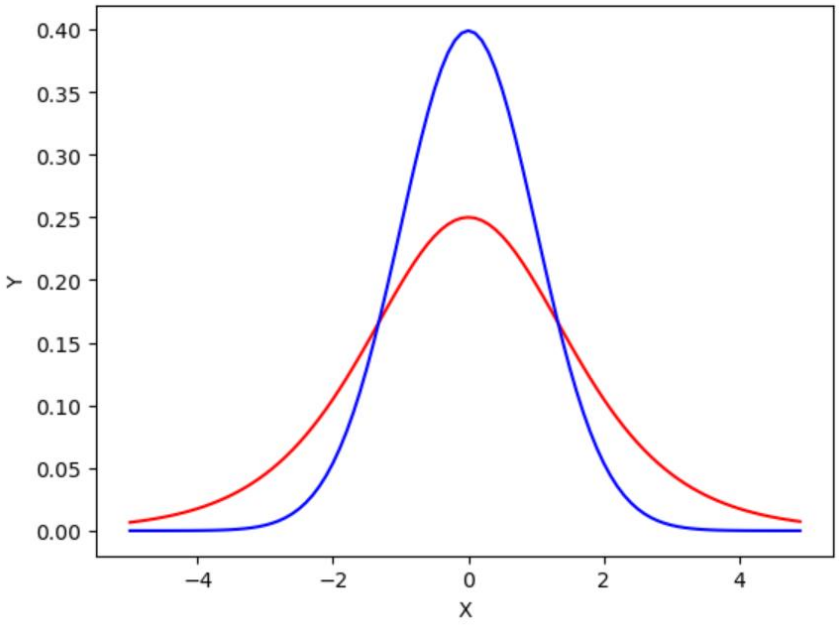

Figure 25: Comparison of the standard logistic distribution (red) and the standard normal distribution (blue).

VIII. BASICS OF MULTIVARIATE ANALYSIS

VIII.1 Introduction.

In the previous chapter we already discussed regression with more than two variables. This is a technique that belongs to multivariate statistical analysis. It very often happens that we have data that describe more than one property of the objects under study. For instance, from a soil sample we can determine the amount of organic matter, the average grainsize, the amount of clay, the density, and several other properties. All these properties may be related to each other. The purpose of many multivariate techniques is to unravel these relations and to discover patterns in the data. This becomes more complicated the more variables we add, and it is less simple than with one or two variables only. With one variable, you can examine statistical distribution parameters, with two you can calculate correlation coefficients that show whether there is a relation between two. In a multivariate dataset, you have a large array of numbers divided over several variables. Then the relations between the measured objects and the variables are more difficult to detect with simple statistics.

As we have discussed in Section VI. ALTERNATIVE SOLUTIONS TO THE REGRESSION PROBLEM: *Reduced major axis (RMA) and principal axis.*, an added complexity is that it is not always easy to separate dependent from independent variables in these datasets. In the soil sample case above, there may be, for example, a correlation between density and organic matter and density and soil acidity (pH). Nice to know, but it does not say anything yet about the true relations between the variables and the processes that cause these relations. Many questions arise from these two simple correlations:

- Is density truly the variable that causes soils to have a higher organic matter content and a lower pH?
- Or is a higher density caused by a higher pH?
- And what about the correlation between density and organic matter?
- Could there be a variable that causes both differences in density and pH?

Some of these questions can be answered using multivariate techniques, but others require you to think logically about the causal relationships in the dataset. In this case it is most likely that the organic matter content determines both density and pH, as soil organic matter contains organic acids and is lighter than mineral matter. This can be shown by a thorough analysis of correlation coefficients, or a technique called **factor analysis**. We will discuss this technique in detail later, but first we need to know a thing or two about multivariate datasets.

VIII.2 Organizing multivariate data: Matrices and data space

Multivariate data come in tables or **matrices**. Usually, in such a matrix a row represents some sample or object of which we have measured several variables. Each column represents one variable. Table 8 shows an example of a matrix for our soil sample example. The rows represent the soil samples. Each entry in one row represents the measurement result of one variable. Row 5 contains the results for sample number V, the first entry in the row is organic matter content, the second average grain size, the third clay content, the fourth density and the fifth shows its pH.

Table 8: Example of a matrix containing multivariate data of soil samples

Sample	Organic matter (%)	Average grain size (mm)	Clay content (%)	Density (g/cm³)	pH
Sample I	3	1.1	21	2.2	6.5
Sample II	8	0.9	22	1.8	5.7
Sample III	12	0.7	35	1.7	5.9
Sample IV	6	1.3	12	1.9	6.1
Sample V	21	0.3	32	0.9	5.8
Sample VI	31	0.2	65	1.2	5.2

Multivariate statistics is based on computations with matrices. You can add or subtract matrices from each other, multiply them and compute the inverse of matrices. We have seen a bit of this in the Extra Reading sections of the Regression chapters above. Further computations with matrices involve computations of determinants and eigen values. This is not repeated here. Part of it already has been treated in previous courses. Davis (2002[1]) contains an excellent primer in matrix algebra. Another good place to start if this is new to you is this page, which goes through the steps of several matrix operations with helpful examples. It is strongly recommended to study the basics of matrix operations, but it will not be a subject of this course or its examination.

A concept that is helpful in understanding multivariate statistical techniques is that of the **data space**. The multivariate observations on an object, for instance the soil samples in Table 8, can be seen as coordinates, defining the location of a point in a space defined by coordinate axes (think about the coordinates on a map). The coordinate axes are the measurement scales of the variables, i.e. the columns of the matrix. For bivariate datasets we have used this concept already many times above. The scatter plot of a bivariate data set is an example in which we represent the observations as points in a plane in which the *x* coordinate represents the value of the first variable and the *y* coordinate the value of the second variable (e.g. Figure 5). This can be extended to more than two variables. Three variables we still can represent graphically in a three-dimensional space (e.g. Figure 16). With four or more variables, we can no longer make a graphical representation of the variable space unless we plot only three or two variables at a time. However, mathematically nothing really changes: A 100-dimensional data space from a dataset with 100 variables requires the same mathematics as a two-dimensional data space.

VIII.3 The correlation matrix

The **variance-covariance matrix** or the **correlation matrix** is usually the starting point for analysis of a matrix containing multivariate data. It contains the variances and covariances of all the variables in the data matrix. Remember from the Statistics part of the course that the population variance of a variable is a measure of the dispersion (or "spread") of the values around their mean, and is computed as:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - \left[\frac{(\sum_{i=1}^{n} x_i)^2}{n}\right]}{n-1} \qquad (54)$$

[1] Davis and Sampson.

The standard deviation (σ) is the square root of the variance. Likewise, the covariance of two variables X_j and X_k denotes the mutual dispersion of two variables around their mean. It is defined as

$$cov_{jk} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{n-1} = \frac{n * \sum_{i=1}^{n} x_{ij}x_{ik} - \sum_{i=1}^{n} x_{ij} * \sum_{i=1}^{n} x_{ik}}{n(n-1)}$$

(55)

Take a closer look at the formula between the two '=' signs in Equation 55. It says that the covariance is computed as the product of the deviations of the values in the variable from the means of the two variables. If these deviations tend to be large throughout the population, the absolute value of the covariance will be large, like the variance. But covariance has an additional property: If for one variable the deviation from the mean is positive while, for the same object or sample, the deviation for another variable is negative, the covariance also will be negative. Likewise, if the variables vary in the same direction, it will be positive. So, the covariance indicates to what extent variables vary together in a positive or negative sense, or in other words are related to each other. We have discussed this effect in Section II. CORRELATION (see Figure 3 and Figure 4 for examples in bivariate datasets).

Covariances may vary strongly according to the measurement units of the variables. This is a disadvantage which hinders comparing covariances. Therefore, they are often scaled by dividing by the variances of both variables:

$$\begin{aligned} r_{jk} &= \frac{covariance_{jk}}{standarddev._j * standarddev._k} \\ &= \frac{cov_{jk}}{\sigma_j * \sigma_k} \\ &= \frac{\sum_{i=1}^{n} x_{ij}x_{ik} - \frac{\left(\sum_{i=1}^{n} x_{ij} * \sum_{i=1}^{n} x_{ik}\right)}{n}}{\sqrt{\frac{\left[\sum_{i=1}^{n} x_{ij}^2 - \left(\sum_{i=1}^{n} x_{ij}\right)^2\right] * \left[\sum_{i=1}^{n} x_{ik}^2 - (\sum_{i=1}^{n} x_{ik})^2\right]}{n}}} \end{aligned}$$

(56)

This is known as the **Pearson's correlation coefficient**, already discussed in Section II. CORRELATION of this course. It varies between -1 and +1, a positive value indicates that both variables vary in the same direction (when one has a larger value, the other also has a larger value), a negative value indicates variation in the opposite direction (when one has a larger value, the other also has a smaller value). Also remember that a **significance test** exists (see Section II. CORRELATION) that tests whether the correlation coefficient significantly differs from 0 (= no relation between the variables).

Using Equation 56 is a lot of work. Imagine calculating all the squares and cross products for a matrix of, say, ten or twenty variables by hand. Luckily, using matrix manipulation it is quite easily done. The necessary sums of squares and cross products in Equation 56 are obtained with only one simple matrix multiplication: multiply the data matrix **X** by its **transpose**:

$$S = X^T * X$$

(57)

To do this, we first take the transpose of **X**, and then multiply it with **X**. Just as a reminder: The

transpose of X is obtained by tilting the matrix X on its side. The rows become the columns and the columns become rows:

$$\begin{pmatrix} 1 & 3 \\ 9 & 5 \\ 1 & 4 \end{pmatrix}^T = \begin{matrix} 1 & 9 & 1 \\ 3 & 5 & 4 \end{matrix}$$

(58)

So, if **X** is a *n*m* matrix (n rows, m columns), $\mathbf{X}^T$ is of size *m*n*. Multiplication of $\mathbf{X}^T$ with **X** yields a *m*m* square matrix **S**:

$$\begin{aligned} S &= X^T * X \\ &= \begin{vmatrix} 1 & 9 & 1 \\ 3 & 5 & 4 \end{vmatrix} * \begin{vmatrix} 1 & 3 \\ 9 & 5 \\ 1 & 4 \end{vmatrix} \\ &= \begin{vmatrix} 83 & 52 \\ 52 & 50 \end{vmatrix} \end{aligned}$$

(59)

On the diagonal of this matrix, we find each column of **X** multiplied with itself. The result is the sum of squares of all elements in the column: The $\sum x_{ij}^2$ terms in Equation 56. The off-diagonal elements contain the sums of the cross-products of the elements in different columns of **X**: The $\sum x_{ij} x_{ik}$ terms. Check the example in Equation 59 to see if you can follow this. If not, check this with the chapter on matrix algebra in Davis (2002[1]) or search for help online.

When we subtract the column averages from every row, the result of Equation 59 is quite similar to the matrix of covariances. The only thing we still have to do to obtain the matrix of covariances from **S**, is dividing **S** by *n-1*, the number of observations minus one. We can make this calculation even easier when we **standardize** the matrix **X**. To do this, we subtract the average of the corresponding column from each element in the matrix and thereafter divide it by the column standard deviation. If $\overline{x_k}$ is the column average and σ_k is the column standard deviation the standardized elements (z scores) are obtained by

$$z_{ik,std} = \frac{x_{ik} - \overline{x_k}}{\sigma_k}$$

(60)

When Equation 57-59 are applied to this standardized data matrix and the result divided by *n-1*, the resulting matrix is the **correlation matrix**. In our very simple example, this would result in:

$$\begin{aligned} S &= {z_{ik,std}}^T * z_{ik,std} \\ &= \begin{vmatrix} -0.58 & 1.15 & -0.58 \\ -1 & 1 & 0 \end{vmatrix} * \begin{vmatrix} -0.58 & -1 \\ 1.15 & 1 \\ -0.58 & 0 \end{vmatrix} \\ &= \begin{vmatrix} 2 & 1.73 \\ 1.73 & 2 \end{vmatrix} \end{aligned}$$

(61)

Standardization is applied very often in multivariate statistics. It has the great advantage that all the variables are expressed on the same measurement scale: units of variance, or z scores (remember the Statistics part of the course). The original measurement scales may represent a wide range of

[1] Davis and Sampson.

numbers, varying over several orders of magnitude. For example, in our soil sample example (Table 8) the pH ranged from 5.2 to 6.5 and clay content from 12% to 65%. In statistical computations this will cause unwanted effects and therefore it is desirable to use the same scale for all variables.

Below follows an example of what information you can read from a correlation matrix. The example dataset is derived from geochemical data of 35 borehole samples from Pleistocene river samples in the Netherlands. The data contains the most important elements, organic matter and clay content of the sediment measured in the lab. The most abundant rock elements (Si, Al, Fe, Mg, Ca, Na, and K) are expressed on an oxide basis, the rarer metals (Cr, Zn and Ni) on an elemental basis. The correlation matrix (Table 9) is symmetrical along the diagonal, because the correlation of column *j* with column *k* is the same as that of column *k* with column *j*. Also, along the diagonal entries, the correlations are always exactly 1 since the correlation of a column with itself is perfect by definition.

Table 9: A correlation matrix of a geochemical dataset with 35 borehole samples in Pleistocene river sediments. Values highlighted in red represent statistically significant correlations ($p < 0.01$). "OM" stands for organic matter content.

	SiO_2	Al_2O_3	Fe_2O_3	MgO	CaO	Na_2O	K_2O	% OM	Cr	Zn	Ni	% clay
SiO_2	1.00	-0.28	**-0.55**	-0.26	-0.43	0.43	0.33	**-0.95**	**-0.60**	-0.43	**-0.56**	-0.28
Al_2O_3	-0.28	1.00	**0.80**	**0.88**	-0.07	0.01	**0.65**	-0.03	**0.84**	**0.81**	**0.82**	**0.66**
Fe_2O_3	**-0.55**	**0.80**	1.00	**0.73**	0.02	-0.19	0.34	0.29	**0.91**	**0.85**	**0.90**	**0.55**
MgO	-0.26	**0.88**	**0.73**	1.00	0.09	0.29	**0.76**	-0.05	**0.77**	**0.66**	**0.81**	**0.47**
CaO	-0.43	-0.07	0.02	0.09	1.00	0.11	-0.14	0.40	0.09	-0.03	0.07	0.15
Na_2O	-0.43	0.01	-0.19	0.29	0.11	1.00	**0.66**	-0.51	**-0.61**	-0.29	-0.10	-0.29
K_2O	0.33	**0.65**	0.34	**0.76**	-0.14	**0.66**	1.00	**-0.58**	0.34	0.34	0.41	0.27
org.	**-0.95**	-0.03	0.29	-0.05	0.40	-0.51	**-0.58**	1.00	0.35	0.20	0.30	0.11
Cr	**-0.60**	**0.84**	**0.91**	**0.77**	0.09	**-0.61**	0.34	0.35	1.00	**0.83**	**0.93**	**0.59**
Zn	-0.43	**0.81**	**0.85**	**0.66**	-0.03	-0.29	0.34	0.20	**0.83**	1.00	**0.87**	**0.71**
Ni	**-0.56**	**0.82**	**0.90**	**0.81**	0.07	-0.10	0.41	0.30	**0.93**	**0.87**	1.00	**0.62**
% clay	-0.28	**0.66**	**0.55**	**0.47**	0.15	-0.29	0.27	0.11	**0.59**	**0.71**	**0.62**	1.00

A significance test for the correlation coefficient is based on the Student's *t*-distribution:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

(62)

Here, *n-2* is the number of degrees of freedom. With some algebra (work this out for yourself to see if you can follow!) this can be converted into a formula that gives the correlation coefficient for any value *t* related to a given significance level:

$$r = \pm\sqrt{\frac{t^2}{t^2 + n - 2}}$$

(63)

For the correlation coefficients in Table 9, the sample number n equals 35, and the value of t belonging to a one-sided significance level of 1% is 2.75. From Equation 63, this corresponds to a correlation coefficient of ±0.44. In the table, all correlations with an absolute value above 0.44 are highlighted in red. These are the correlations in our dataset which are statistically significant with a significance level of 1%.

From this, a pattern of high correlation emerges: the clay percentage, the metals Cr, Zn and Ni, and Al_2O_3, Fe_2O_3 and MgO show high mutual correlations. SiO_2 (the main constituent of sand) correlates negatively with most of this group. For anyone with some knowledge of mineralogy and geochemistry this is not surprising. Clay minerals consist largely of aluminum oxides, and the clays in this area are rich in iron and magnesium (smectite clays). Metal ions such as zinc, chromium and nickel easily adsorb to clays. So, we can conclude that the chemistry of all the borehole samples is largely determined by the clay content.

This example illustrates nicely that an analysis of the correlation coefficients of multivariate data can already enhance your insight into the causes of the variation in the data. In a later chapter we will discuss factor analysis, which can help to clarify more of these kinds of patterns in a dataset, but first we will dive a bit deeper into multivariate correlations.

VIII.4 Induced correlations and closed datasets

In some cases, high correlations can be artefacts caused by computational procedures: **induced correlations** (see Davis, 2002[1]; this blog post contains a helpful example). An important source of erroneous correlations are **closed datasets**, which unfortunately occur very often in Earth sciences.

In a closed dataset, all variables measured on the objects of the sample population add up to a fixed number, for instance 1 or 100%. This often happens with data on composition of samples. For instance, gravel samples from a riverbed may be analyzed on the type of rock that each gravel grain is composed of. Usually, this composition is expressed as a percentage or fraction. For example, a sample consists of 45% of quartz grains, 33% of sandstone, 12% igneous rocks and 10% of limestone. The percentages add up to 100% and will do so for every gravel sample that is taken from the river. Now imagine that you have taken another sample of which, after counting the grains, you conclude that it contains 68% quartz. Automatically, the percentages of other components (sandstone, igneous rocks, and limestone) must be lower. This causes correlations between these variables (sandstone, igneous rocks, and limestone content) with the variable describing the quartz content to be more strongly negative than they should be when no percentages were calculated. This phenomenon is sometimes referred to as the **closed sum effect**.

To illustrate this, two correlation matrices are shown in Table 10. The leftmost one is from a data matrix of 100 x 4 elements (100 measurements of 4 variables). All of them were generated by a random number generator. As expected from random numbers, there should be no correlation between the columns of the matrix. Indeed, in the leftmost matrix the correlations are all close to zero.

The rightmost correlation matrix is also derived from the same data matrix. However, now the rows

[1] Davis and Sampson.

of the matrix have been added up, and each row element has been expressed as a fraction of the total. For clarity, here is the sum of the first row of the data matrix:

$$0.9501 + 0.5828 + 0.4398 + 0.3603 = 2.3330 \quad (64)$$

The resulting fraction-of-total values of this row used in the correlation matrix on the right of Table 10 is calculated as follows:

$$\left|\frac{0.9501}{2.333} \quad \frac{0.5828}{2.333} \quad \frac{0.4398}{2.333} \quad \frac{0.3603}{2.333}\right| \rightarrow |0.4073 \quad 0.2498 \quad 0.1885 \quad 0.1544| \quad (65)$$

What is immediately striking in the rightmost correlation matrix in Table 10 is that all the correlations are now negative and have considerably higher absolute values than in the leftmost matrix. Still, these correlations are derived from the same random, originally uncorrelated data! With a real data set, e.g. the river gravel composition, this could suggest relations between the different rock types that are not there.

Table 10: Example of two correlation matrices for the same dataset with 100 samples of 4 variables. For the matrix on the left, the variables are left untreated. For the matrix on the right, the values in each row are recalculated as fractions of a total (1).

correlation original random data					***correlation random data expressed as fraction of row total***				
variable	**A**	**B**	**C**	**D**	**variable**	**A**	**B**	**C**	**D**
A	1.000	-0.066	-0.037	0.022	**A**	1.000	-0.302	-0.369	-0.248
B	-0.066	1.000	-0.124	-0.062	**B**	-0.302	1.000	-0.373	-0.356
C	-0.037	-0.124	1.000	-0.115	**C**	-0.369	-0.373	1.000	-0.346
D	0.022	-0.062	-0.115	1.000	**D**	-0.248	-0.356	-0.346	1.000

With compositional data it is unfortunately impossible to say how much of the correlations are induced and how much is due to real relations between the variables. A way to avoid this problem of induced correlation is to use Aitchison's **log-ratio transformation**. The procedure is as follows:

1. Select a variable that is non-zero for all rows of the data matrix. The selected variable is denoted by *s*.
2. Divide the values of the other variables by the value of the variable *s* in the corresponding row, forming ratios between all variables and *s*: For row *i*, the ratio $x_{ij,ratio} = \frac{x_{ij}}{x_{is}}$.
3. Next take the logarithm of all ratios: $x_{ij,transformed} = \ln\left(x_{ij,ratio}\right) = \ln\left(\frac{x_{ij}}{x_{is}}\right)$

The resulting transformed variables can vary freely between -∞ and +∞. The calculation of covariances from the transformed values proceeds in a different and more intricate way than the ordinary correlation matrix in the previous section. It will not be discussed here but Davis (2002[1])

[1] Davis and Sampson.

gives a more extensive treatment.

A simpler approach to obtain meaningful covariances is the **centered log-ratio covariance**. This is obtained by:

1. Take logarithms of all elements of the data matrix.
2. Determine the averages of the logarithms of every row of the data matrix.
3. Subtract these averages from each element of the row.
4. Calculate the covariances from the data matrix transformed in this way in the usual manner outlined in the previous chapter.

A disadvantage of these procedures is that they do not work well when there are zero values in the data or missing values, because the logarithm of zero is not defined. Another disadvantage of the log-ratio transformation is that we lose information from one of our variables since we use that variable (*s*) as the base of our ratios. Finally, another disadvantage of log-transforming our data is that the distribution of the data is affected, because the logarithmic operation affects values close to zero differently from values further away from zero. This has implications for further data treatment, such as significance testing, for which a normal distribution of the data is often assumed (remember the explanation of distributions in the Statistics part of the course)

VIII.4 Take Home Messages

- Multivariate datasets are datasets with more than 2 variables.
- Multivariate data is generally stored in a **matrix**, in which the rows represent objects or samples, and the columns represent variables.
- A **correlation matrix** gives us a useful first impression of the covariation and correlations between the variables in our multivariate dataset.
- In **closed datasets**, in which all values for a sample sum to a constant (e.g. 100%), **induced correlations** may occur, which can bias our interpretation of the data.
- Transformations such as the **log-ratio transformation** or the **centered log-ratio covariance** can help us derive meaningful information about the covariance within closed datasets.

IX. MULTIVARIATE DISTRIBUTIONS AND CLASSIFICATION

IX.1 The multivariate normal distribution.

In the first part of this course the normal or Gaussian probability distribution has been presented. This distribution is used for a single variable. However, when we have two or more variables, we can also draw up a normal probability distribution that takes account of the variance and covariance of both variables.

Suppose we have two variables X_1 and X_2. The populations of all values of these variables are normally distributed. So if you want to calculate the probability that a certain value x_1 is part of the population of X_1, you can use the normal distribution of X_1 with its parameters μ_1 (average) and σ_1 (standard deviation). The same holds for X_2. However, when our population consists of objects of which the two variables X_1 and X_2 are measured attributes (instead of single, unconnected variables), it is better to use the multivariate normal distribution. This is especially true when the variables are correlated. Why this correlation matters will be explained on an example later in this section. First, let us go through some theory and an explanation on what the multivariate normal distribution looks like.

We start with a two-variable example. When you have measured two variables on a population of objects, such as the mean annual temperature and the amount of precipitation at a series of locations, you can make separate histograms and a scatter plot of all temperatures and all precipitation values (Figure 26). However, it is also possible to construct a histogram of the observations in two dimensions, as shown in Figure 27A.

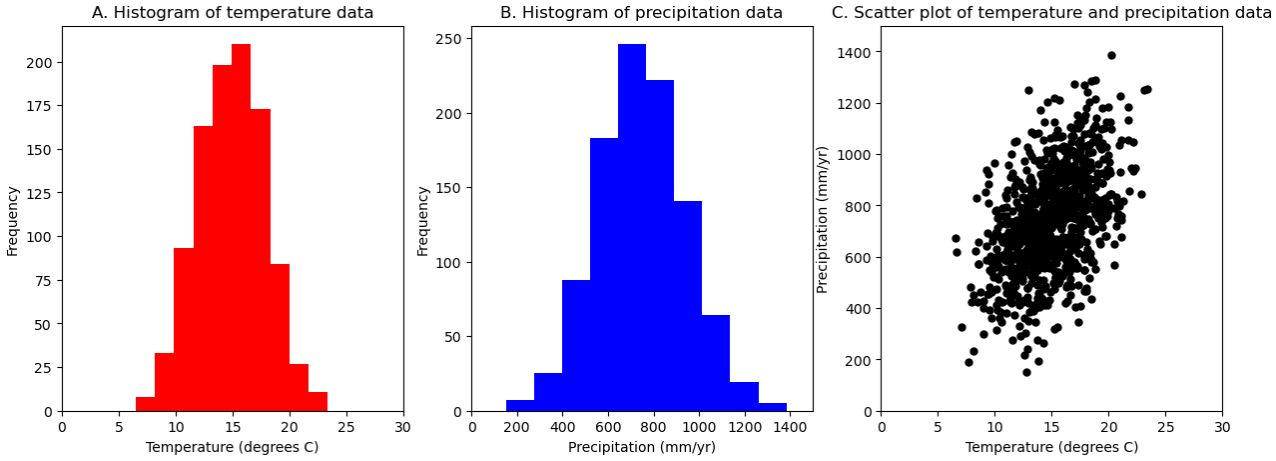

Figure 26: Individual histograms and a scatter plot of the variables temperature and precipitation measured at several locations.

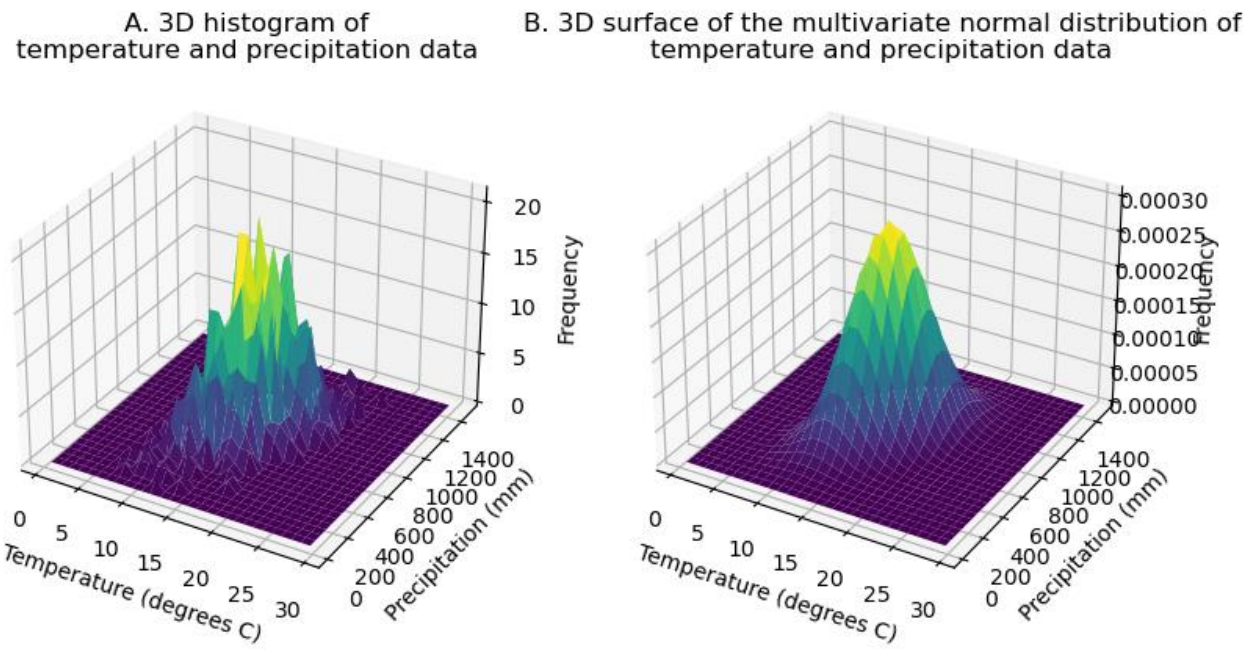

Figure 27: A) 3D plot of a histogram of 1000 samples taken from the multivariate normal distribution of the variables temperature and precipitation. B) Smooth 3D-plot of the underlying multivariate normal distribution of the variables temperature and precipitation measured at several locations. The correlation between the two variables is 0.5.

In Figure 26 the right panel shows a scatter plot of 1000 observations of two correlated variables, temperature and precipitation. Just like in the single variable case, the axis of temperature and precipitation can be subdivided in equal-sized classes. Based on that, you could construct two histograms of both variables separately (Figure 26, left panels). However, since the variables are correlated, we should incorporate the correlation between the two variables into the graphs as well. So we count the joint values of both variables. This is done by dividing the sample space of temperature and precipitation in equal sized rectangles and counting the number of observations in each rectangle.

Figure 27A right shows a frequency surface which was constructed from these counts. This is an analogue of the frequency polygon or histogram in the single variable case. The vertical axis is the number of counts (frequency). Note that the peaks of this surface reflect the density of points in Figure 27B. The latter shows the true multivariate normal distribution from which the dataset in Figure 27A was sampled. The sampling is imperfect (even though we took 1000 samples), and this causes the rough "spikey" pattern on top of the histogram surface. If we would sample an infinitely large number of times, we would get a surface that approximates the one in Figure 27B.

Like in a univariate case (see Statistics part of the course), we can fit a probability density function to these frequency counts. This seems a bit silly, because we already used the distribution plotted in Figure 27B to sample the data plotted in Figure 27A, but remember that, if this was a geological or climatological dataset, we would have the samples but the underlying distribution is unknown. For the normal distribution, a multi-variable version exists. For demonstration (not for memorizing) the formulas for this distribution are given below.

$$f(x) = \frac{1}{2\pi\sigma_1\sigma_2(1-\rho^2)} e^{\left[\frac{-\left\{\frac{(x_1-\mu_1)^2}{{\sigma_1}^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{{\sigma_2}^2}\right\}}{2(1-\rho^2)}\right]}$$

(66)

Equation 66 is the version for the bivariate (two variable) case. The formula is similar to the formula for the normal distribution (see Equation 53) and it also contains the population means (μ_1, μ_2) and variances (σ_1, σ_2) of the two variables. In addition, it also contains the correlation coefficient, ρ.

$$f(x) = \frac{1}{(2\pi)^{N/2}\sqrt{|R|}} e^{\left[\frac{-(x-\mu)^T R^{-1}(x-\mu)}{2}\right]}$$

(67)

Equation 67 is the multi-variate version and Equation 66 for two variables can be derived from it. In Equation 66, *N* is the number of variables, **R** the covariance matrix of all variables. **|R|** is the determinant of **R**. A determinant is a single number computed from a square matrix (see the matrix algebra chapter in Davis, 2002[1]). The *x* and μ in Equation 67 are vectors with *x* representing a vector with a single observation for each variable, and μ containing the population means of each variable. 'T' means the transpose of a vector or matrix, $\mathbf{R}^{-1}$ is the inverse of **R** (see Davis, 2002).

IX.2 Confidence ellipses

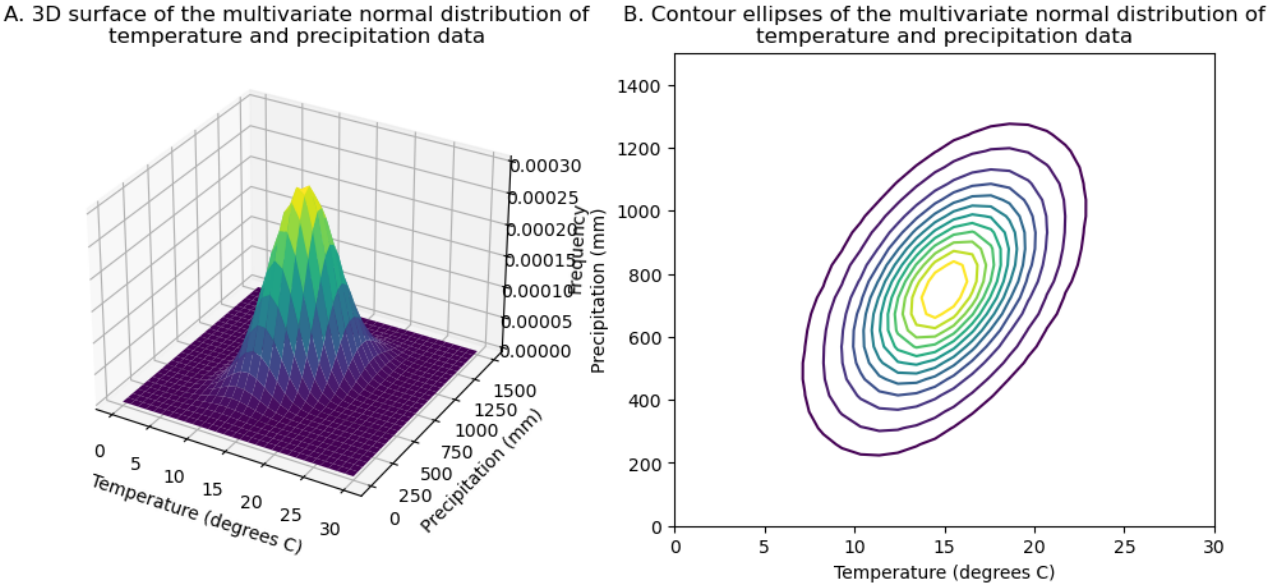

Figure 28: Figure (A) showing a 3D-plot of a multivariate normal distribution of the variables temperature and precipitation and (B) showing contour lines of this distribution.

Figure 28A shows the multivariate normal probability distribution derived from the data of Figure 27A. If you make a vertical cut in any direction of the horizontal plane through 'hump' in the surface representing the probability density function, you will see a single variable normal distribution (as in Figure 26). The horizontal cross section of the function is elliptical, as you can see from the contour lines of the surface (Figure 28B). Every contour line in Figure 28B represents a confidence ellipse, which outlines an area in the X_1-X_2 (in this case temperature-precipitation) plane within with a certain cumulative probability. Confidence ellipses can be constructed for every probability level. We will now consider these ellipses more closely and examine their relationship with correlation of the data. *This is crucial to understand the techniques that are discussed in the following chapter.*

The confidence ellipses have a major (long) and a minor (short) axis. The major axis is also known as the *principal axis* - the same principal axis that was discussed in VI.4: Principle Axis Regression. As you can see from Figure 28B, this principal axis is not aligned to one of the axes. This is always the

[1] Davis and Sampson.

case for correlated variables. If the two variables were uncorrelated (Pearson's r = 0), the principal axis would be parallel either to the axis of variable 1 (e.g. temperature, so horizontal in Figure 28B) or axis of variable 2 (e.g. precipitation, so vertical in the example in Figure 28B), whichever variable has the largest variance. Remember that the 'point cloud' in a scatter plot of two uncorrelated variables stretches parallel to one of the axes.

In the case of correlated variables, as in Figure 28B, the variance along axes of each of the variables is not the largest variance that can be found in the data set. The spread of the data points along the principal axis is much larger, which can be readily seen in Figure 28B. The minor axis, perpendicular to the principal axis, represents the smallest variance in the data set. We will return to this when we discuss Factor Analysis. First, we will discuss another application to of the multivariate normal distribution.

IX.3 Introduction to clustering and classification

Classification is a common problem in data analysis. Classification assigns an individual object or sample to a group of objects with similar characteristics. Classification always requires the development of criteria on which this assignment can be based with the least amount of ambiguity. After all, we need a way to determine which groups is the best fit for each sample. Later, we will discuss some ways in which we can do this. But first let's consider an example derived from remote sensing, a discipline in which classification is very common:

Satellite images are often used to make maps. Satellites usually make measurements of the light intensity reflected from the earth surface in different parts of the electromagnetic spectrum. They combine this data for a small part of the earth surface, effectively cutting up the surface into small area. This results in an image that consists of a regular grid of *pixels* (picture elements). For each pixel, the intensity value for different parts of the electromagnetic spectrum (different "colors" of light) is recorded. For example, each pixel stores values for the intensity for blue light, green light, red light, and very often also for different infrared or even ultraviolet wavelengths.

The various parts of earth's surface have different reflectance characteristics. This is why we observe colors: green vegetation reflects green light strongly, and hardly any red. Most rocks hardly reflect green but reflect more red light. The ocean absorbs a lot of energy and radiates very little back, which is why it looks dark blue in visible light and appears dark in the infrared wavelengths (which carry the energy we experience as warmth). Using this type of information, we can classify surface types from the reflectances associated with the satellite image pixels.

IX.4 Box classification

Suppose we have collected the reflectance values of several satellite image pixels in two parts of the light spectrum. From all these pixels it is exactly known which class of surface types the represent, e.g. a surface type A, B and C. A scatter plot of all the observations is constructed (see Figure 29). The different surface types stand out as clusters of points in the plot. Each cluster represents a different population of pixels, with different mean values, standard deviations, and correlation coefficient for the two variables.

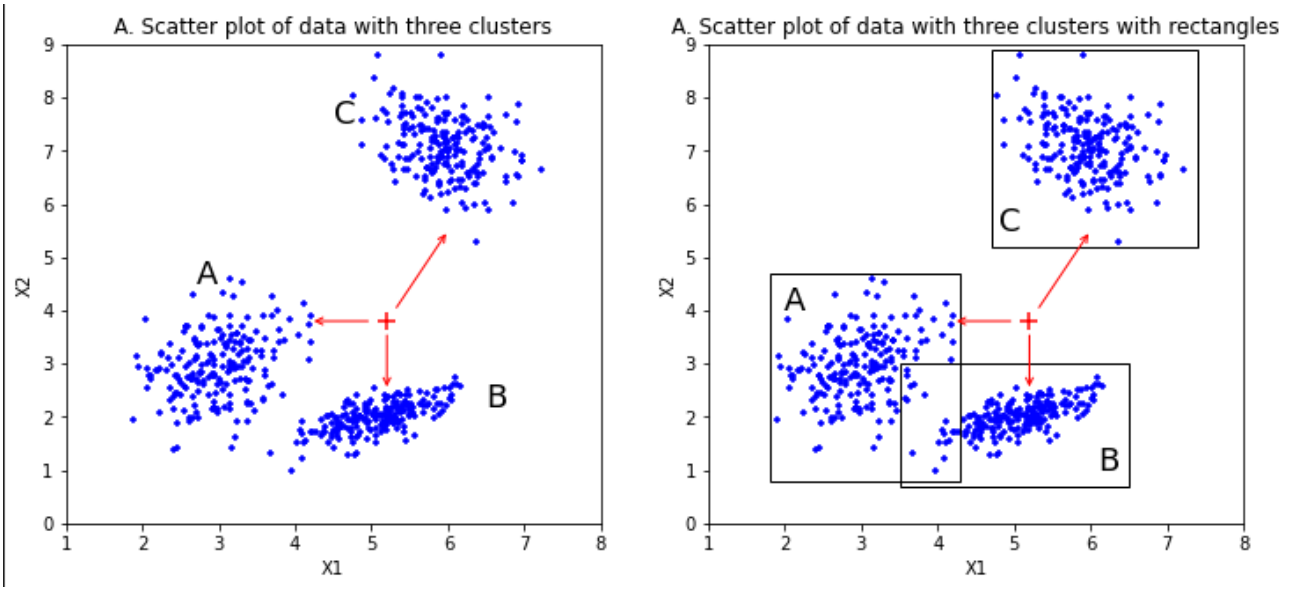

Figure 29: Classification with two variables. Left: observations plotted in a scatter plot. Different groups of observations are discernable by clusters of points. The red point is a data point which should be assigned (classified) to one of the groups. Right: the values of the variables for different groups overlap, which makes classification on value ranges difficult.

The next step is to classify all unknown pixels in the satellite image. If all pixels are classified this results in a map, showing the classes of surface types, e.g. a land use map. The simplest way to do this is taking the reflectance values (values of the variables X_1 and X_2) and see if the values of an unknown pixel fit into the value ranges of the known pixel groups. With approach, we consider the value of the unknown pixel (colored red in Figure 29) and check whether the pixel falls within the range of one unique group. In Figure 29A we can see that the X_1 value of the unknown point falls within the spread of the data X_1 data of class B and class C. However, it's X_2 value is higher than any pixel in class B and lower than any point in class C, so this does not work. The unknown point can also be classified based on the X_2 value in class A. However, it falls outside the range of X_1 values for class A.

We can easily visualize this problem by drawing rectangles around the classes (Figure 29B). This simple '**box classification**' method does not work very well, and it will be even more difficult to perform when more than two variables are involved. The problem is that, in most datasets, the value ranges of different classes overlap for at least one of the variables, and we have no objective way to decide which variable we should "trust" better than the other.

IX. 5 Maximum Likelihood Classification

A better and more objective approach is the **maximum likelihood classification**, which is sometimes (especially in the Remote Sensing literature) referred to as **supervised classification**. This is very often applied in image processing. It consists of the following steps:

1. Determine the multivariate normal population parameters of all the groups of known objects, using a sample of their populations. In our example of Figure 29, the parameters of the three clusters A, B and C will be determined as separate populations. The population distributions are shown in Figure 30.
2. For every unknown object, compute the probability that it belongs to one of the populations based on their respective probability distributions. In our example this would result in three probabilities for the red point P: $p(P \in A), p(P \in B), p(P \in C)$ These are the probabilities that P belongs to A, B or C.

3. Assign the point to the population with the highest probability.

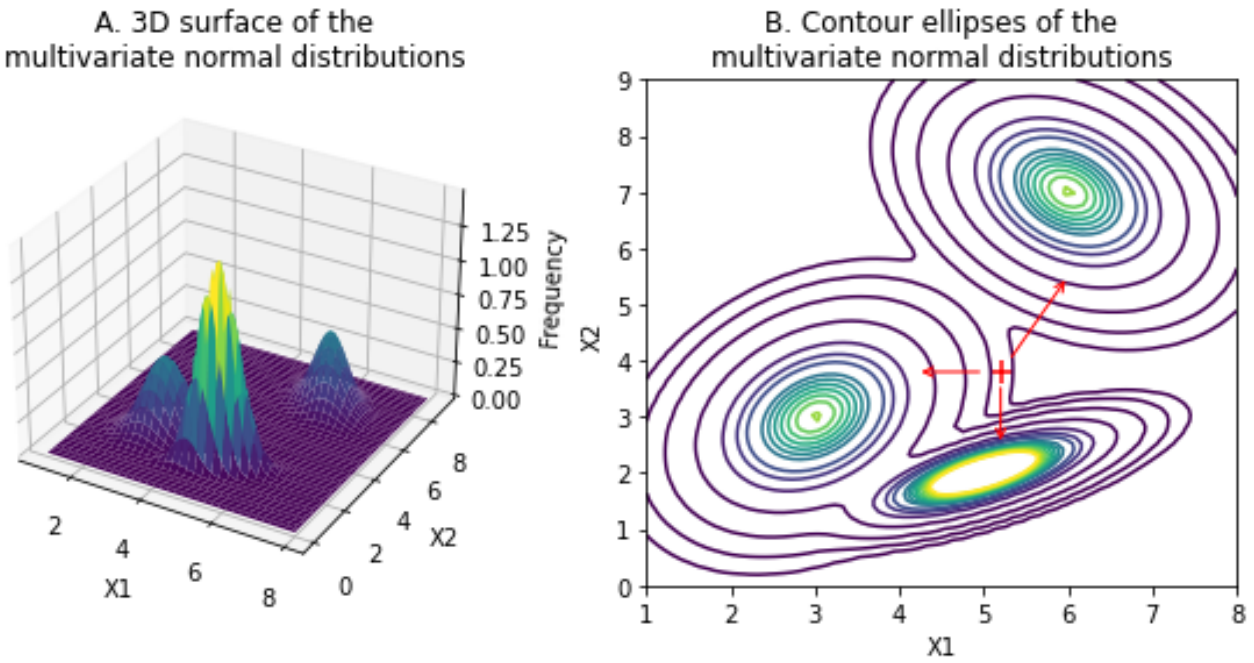

Figure 30: A. Multivariate normal probability density functions of the sample groups A, B and C in Figure 29. B. The same probability density functions shown as contour plots.

The contour lines plot in Figure 30B also nicely illustrates how this works: Each contour represents the same probability, so the contours "spread out" from the center of each class. The class which contour "hits" the unknown datapoint first is the most likely class to which the datapoint should belong. In this case, that is class A. Perhaps this surprises you because the center of class B is closer to the location unknown point, but class A has a wider distribution (with a higher standard deviation) and therefore it is more "welcoming" to new datapoints.

This is a very short outline of the maximum likelihood classification without mathematical details. However, it outlines very well the use of the multivariate normal distribution as well as the process of classification. If you want to learn more about maximum likelihood classification and see some useful examples of how it works on actual satellite data, this page by the GIS (Global Information Systems) software producer ESRI can be useful. Should you want to dive into the mathematics behind this classification method, this paper[1] gives a good overview with examples from the field of psychology.

IX.6 Hierarchical Clustering

Maximum likelihood classification is just one of many ways in which you can classify or cluster samples into groups. It is ideal for research questions where the main aim is to assign each datapoint to the most likely or nearest cluster, and where the distance to the other clusters or classes does not matter so much. However, there are lots of problems in research where the *order* in which the classes fit the unknown datapoints matters, or where we are interested in the *degree of similarity* between datapoints.

A clear example of this is when we want to construct a tree of life (called a "phylogeny" in the biological literature). A tree of life, or a family tree, does not only contain information about the

[1] In Jae Myung, "Tutorial on Maximum Likelihood Estimation," *Journal of Mathematical Psychology* 47, no. 1 (February 1, 2003): 90–100, https://doi.org/10.1016/S0022-2496(02)00028-7.

classes or groups in which the members are clustered, but also about how different the datapoints and groups are to each other. In other words, in this type of classification the *hierarchy* of the datapoints matters, which is why we call this technique **hierarchical clustering**.

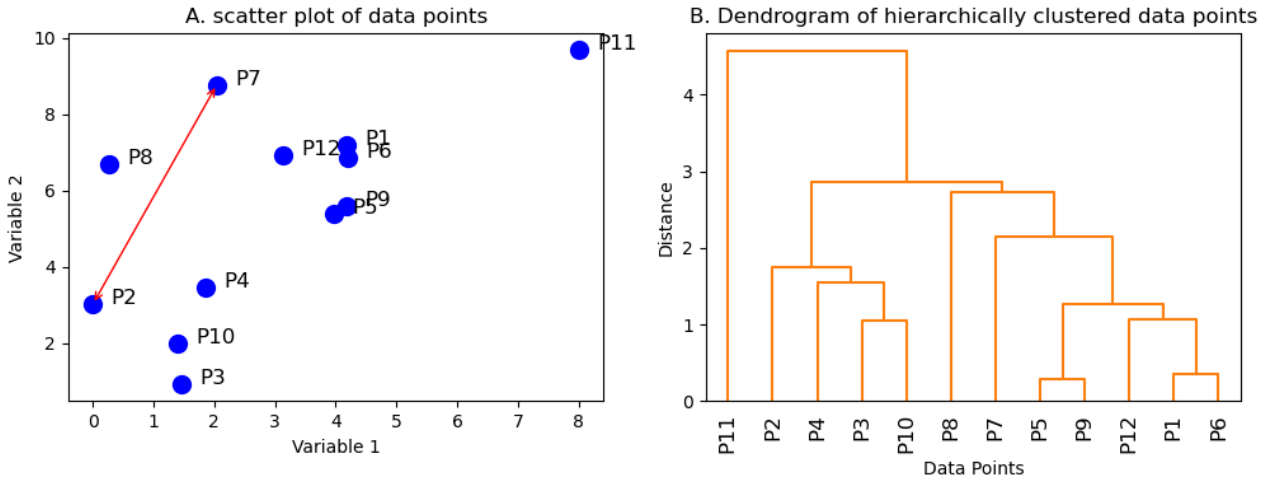

Figure 31: Example of hierarchical clustering of 12 datapoints (P1-P12) with 2 variables. A. shows a scatter plot of the data in the direction of the two variables. The red arrow illustrates the Euclidian distance between P2 and P7. B. shows the dendrogram resulting from hierarchical clustering of the data.

An example of hierarchical clustering is provided in Figure 31. In Figure 31A, we see a scatterplot of a set of 12 datapoints of which 2 variables are measured. As you can see, some points (such as P5 + P9 and P1 + P6) are very close together while other points (such as P11) are far removed from the rest. If we want to cluster these points without losing the relative differences between them, we need to calculate for all the possible combinations of points what the **similarity** is between them. In this example, we will use the **Euclidian distance** as a measure for the similarity (see Figure 31A). The Euclidian distance is literally defined as the distance between the two points in the multi-dimensional space described by the parameters. In this case, with only two parameters, it can be thought of as the length of a straight line that connects two points together. As you probably know, we can use the Pythagorean Theorem to calculate this distance:

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}$$

(68)

In this function, the values p_1 and q_1 are the values for Variable 1 of point p and point q, while p_2 and q_2 are the values for Variable 2 of point p and point q, respectively. For multiple variables we can simply expand this function as follows:

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots (p_n - q_n)^2}$$

(69)

Here, n is the total number of variables in our dataset. Euclidian distance is not the only way to determine the similarity between two points in a dataset though, and we will discuss some other options in IX.7 Similarity Metrics below.

The points which have the smallest Euclidian distance are most similar (in this case P5 and P9). They are clustered together to form the first cluster. We then repeat this process again, treating P5 and P9 together as one point (for example by taking the averages for both variables for these two points) and find the next cluster. If we continue to do this, we are left with an ever-decreasing number of points, until we are left with one big cluster. The hierarchical clustering algorithm can be summarized

as follows:

1. Calculate the similarity between all points in the dataset.
2. Find the two points with the highest similarity (or lowest difference).
3. Cluster these two points by calculating the mean value for the points for each variable.
4. Replace the original two points with the new, combined datapoint.
5. Repeat step 1.

If we create a plot which shows the names of the original datapoints on the horizontal axis and the similarity between the points on the vertical axis, we can easily visualize this process of hierarchical clustering by connecting groups of points which end up in the same cluster with horizontal lines at the height of their similarity. A plot like this is called a **dendrogram** (or "tree diagram", dendro = tree in Greek), and it is shown for the example dataset in Figure 31B. If we read the dendrogram from the bottom to the top, we can follow the successive clustering steps in the algorithm. Note how this dendrogram looks very much like an upside-down family tree, with two datapoints (or persons, in the case of a family) leading to a new datapoint (cluster in statistical terminology). A family tree is basically a dendrogram, with the major difference being that it is not possible to have siblings (brothers and sisters, or multiple new datapoints coming from one pair) using the clustering algorithm we outlined above. There are variants of hierarchical clustering that do allow this though. We will not discuss those here.

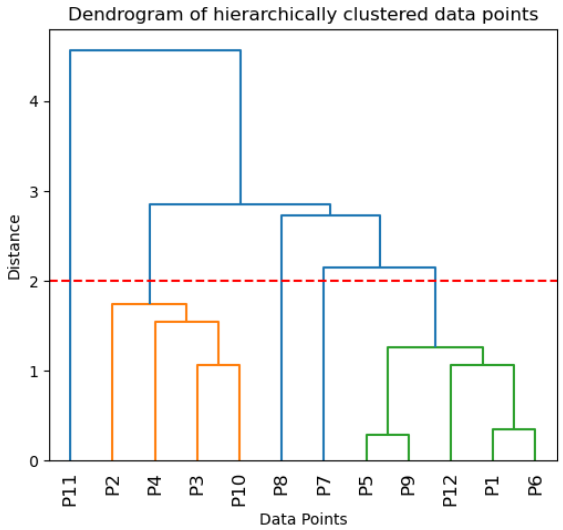

Figure 32: Dendrogram showing the result of hierarchical clustering of 12 datapoints (P1-P12) with 2 variables and with a cutoff line at a Euclidian distance of 2 (dashed red line).

So how do we use this dendrogram to make clusters? To do this, we need to determine which similarity we consider acceptable for forming clusters. Basically, what we need to do is draw a horizontal *cutoff line* on our dendrogram. All points that are connected below that line are considered to be within the same cluster, while all points whose connection is "cut" by the line belong in different clusters. Figure 32 shows how this works for the same dendrogram as shown in Figure 31B. Note that the cutoff at a Euclidian distance of 2 forms two clusters: Cluster 1 contains the points P2, P3, P4 and P10 while Cluster 2 contains P1, P5, P6, P9 and P12. The remaining points are too dissimilar to belong to a cluster. Note, however, that this choice of a cutoff line of 2 is somehow arbitrary. We could have chosen a slightly higher cutoff (moving the red line upwards), which would include P7 in the green cluster, or we could have been stricter about our choice of

cutoff value, which might have left P2 outside the orange cluster. As always, peep in mind that these types of choices are made by the researcher and should be motivated by your research question. No computer algorithm can determine these choices for you, showing again that data analysis is a tool in your research and that it does not do the research for you!

IX.7 Similarity Metrics

In the example above, we had to start by calculating the similarity between each set of points. The method we choose can have consequences for our result! In the example above, we have used the **Euclidian distance** between the points as a measure for their similarity. The method we use to determine the similarity between two points in a clustering dataset is called a **similarity metric**, and there are many types of similarity metrics you can use!

The best type of similarity metric to use is highly dependent on the type of data we have in our dataset. Hierarchical clustering is a very important operation in data analysis, and it is applied on many types of data. As in regression problems, we can deal with continuous, discrete or binary datasets. As you learned in the Statistics part of this course, not all datasets have variables that are normally distributed. All these differences can influence your choice of which similarity metric to use in clustering. We will briefly discuss two simple options for continuous datasets (like the one in Figure 31) in more detail here, but if you are interested in other possibilities, this blog post explains 17 different similarity metrics with clear illustrations. You might encounter some more exotic ones later in your career!

A first alternative to the Euclidian distance is the **Manhattan distance**. The Manhattan distance, also sometimes referred to as the "cityblock distance", calculates the distance between two points in the dataset along the axes of the variables. In other words, it calculates the distance using only right angles (Figure 33). This is why it is called the "Manhattan distance", because it resembles the route you have to take when getting from one place to another in Manhattan (or any other city with a grid-like street pattern). The Manhattan distance can be calculated by taking the sum of absolute distances between two datapoints for every variable:

$$d(p,q) = \sqrt{|p_1 - q_1| + |p_2 - q_2| + \cdots |p_n - q_n|} \quad (70)$$

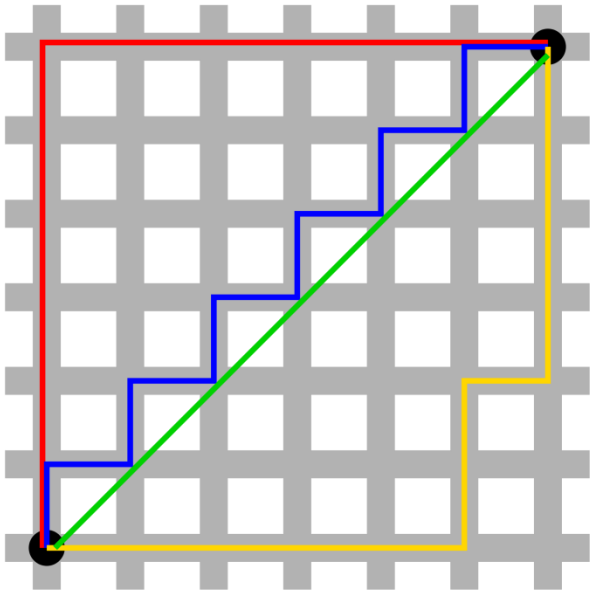

Figure 33: Graphical comparison between the Euclidian distance (green) and the Manhattan distance (other colors).

Another similarity metric is already familiar to you: the **Pearson's r**, or Pearson's correlation distance. We have encountered in II. CORRELATION and the formula for calculating it is:

$$r(p,q) = \frac{cov(p,q)}{\sigma_p * \sigma_q} = \frac{\sum(p_i - \bar{p}) * (q_i - \bar{q})}{\sqrt{\sum(p_i - \bar{p})^2 * \sum(q_i - \bar{q})^2}}$$

(71)

In this formula, *p* and *q* are two points in the dataset defined by values in *i* variables. $\bar{p}$ and $\bar{q}$ are the average values of the points in all variables. This only works if the values within all variables in the dataset are on the same scale. The best way to achieve this is by *standardizing* the dataset (see VIII. BASICS OF MULTIVARIATE ANALYSIS).

Figure 34 shows the dendrograms of a hierarchical clustering for the same dataset using different similarity metrics. Note that, in this case, Euclidian distance and Manhattan distance yield the same clustering (although their dendrograms are slightly different), while the Pearson's yields a completely different result. It seems that Pearson's r was not the right choice for this dataset, most likely because we only have two variables and the correlation between two sets of two variables is always perfect (r = 1).

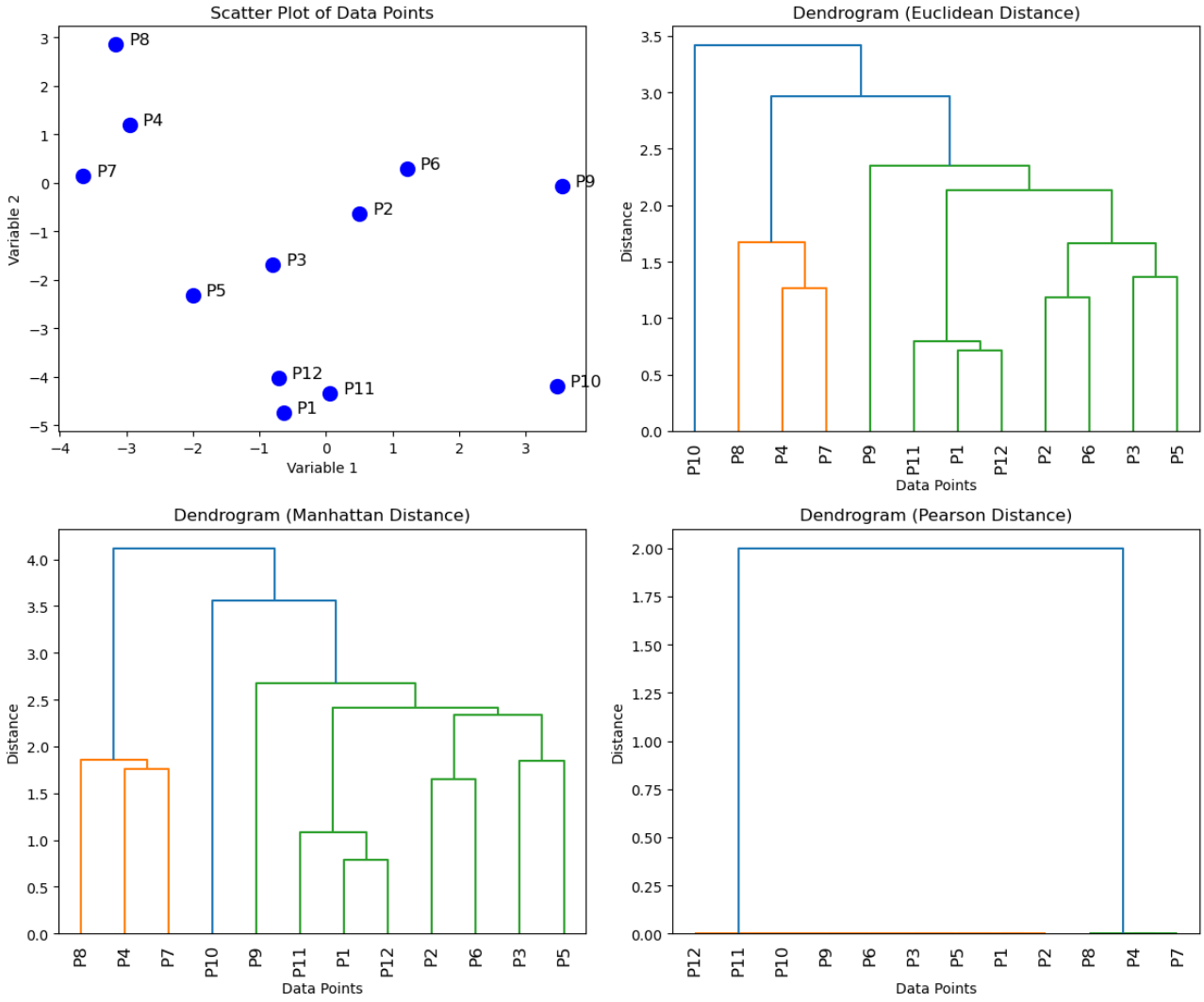

Figure 34: Dendrograms showing the result of hierarchical clustering of 12 datapoints (P1-P12) with 2 variables (scatter plot in top left) and using different similarity metrics (top right and bottom right and left).

IX.7 Take Home Messages

- A **multivariate distribution** is defined in multiple dimensions.
- In a **multivariate normal distribution**, the distribution in the direction of each variable is normal and the multivariate distribution is defined by the distributions with respect to each variable as well as by the **covariance** between the variables
- We can use multivariate distributions to classify datapoints according to their likelihood (or probability) of belonging to different groups within the dataset. These groups (or classes) are defined by multivariate distributions and this type of classification is called **Maximum Likelihood Classification**.
- **Hierarchical classification**, or hierarchical clustering, takes into account the relative difference between samples in the dataset (their "hierarchy").
- We can visualize a hierarchical clustering result using a **dendrogram** and then use a **cutoff value** to determine at which level we want to separate our clusters.
- We can use several **similarity metrics** as a basis for hierarchical clustering. Examples include (but are not limited to!) **Euclidian distance**, **Manhattan distance** and **Pearsons r**.

X. FACTOR ANALYSIS

X.1 The theory behind factor analysis: Loadings and Scores

Factor analysis is a group of techniques that aim to discover a simple, underlying structure in multivariate data. It assumes that behind the different variables in a multivariate dataset, with their many different relations between the variables, one or few **factors** exist that determine these relations. Knowledge of these factors has two major advantages:

1. We may develop a better understanding of the data and the processes that generated them.
2. We can simplify the data by reducing the number of variables or the 'dimensions' of the data (data reduction).

Factor analysis is computationally complex and requires a computer. Moreover, many techniques with confusing names exists. In this syllabus, we will discuss only two simple techniques, starting with the simplest: **principal component analysis**, or **PCA**. Before embarking on the mathematics (which will be kept as simple as possible), it is important to understand the ideas behind factor analysis.

***Disclaimer**: In most years, Factor Analysis is considered the most difficult subject in our Data Analysis course. We do not expect you to know all the ins and outs of the technique at the end of the course.. The goal is for you to grasp the basic principles behind factor analysis and, to be able to use it on simple cases and to interpret its results. Do not worry if you have trouble understanding the concepts explained here at first, this is normal and shows that you are paying attention!*

To show what is meant with **factors** that determine the relations between variables, we return to the correlation matrix shown in section VIII.3 The correlation matrix (repeated below in Table 11). The analysis data contain the most important elements, organic matter and clay content of sediment samples from a borehole. We have removed the values in the bottom left half of the table for simplicity. Remember: These values are by definition the same as those in the top right half of the table.

Table 11: A correlation matrix of a geochemical dataset with 35 borehole samples in Pleistocene river sediments. Values highlighted in red represent statistically significant correlations ($p < 0.01$). "OM" stands for organic matter content.

	SiO_2	Al_2O_3	Fe_2O_3	MgO	CaO	Na_2O	K_2O	% OM	Cr	Zn	Ni	% clay
SiO_2	1.00	-0.28	**-0.55**	-0.26	-0.43	0.43	0.33	**-0.95**	**-0.60**	-0.43	**-0.56**	-0.28
Al_2O_3		1.00	**0.80**	**0.88**	-0.07	0.01	**0.65**	-0.03	**0.84**	**0.81**	**0.82**	**0.66**
Fe_2O_3			1.00	**0.73**	0.02	-0.19	0.34	0.29	**0.91**	**0.85**	**0.90**	**0.55**
MgO				1.00	0.09	0.29	**0.76**	-0.05	**0.77**	**0.66**	**0.81**	**0.47**
CaO					1.00	0.11	-0.14	0.40	0.09	-0.03	0.07	0.15
Na_2O						1.00	**0.66**	-0.51	**-0.61**	-0.29	-0.10	-0.29
K_2O							1.00	**-0.58**	0.34	0.34	0.41	0.27
org.								1.00	0.35	0.20	0.30	0.11

	SiO_2	Al_2O_3	Fe_2O_3	MgO	CaO	Na_2O	K_2O	% OM	Cr	Zn	Ni	% clay
Cr									1.00	**0.83**	**0.93**	**0.59**
Zn										1.00	**0.87**	**0.71**
Ni											1.00	**0.62**
% clay												1.00

From the pattern of high correlations, and with some geochemical background knowledge, we could deduce that the clay content of the samples should be responsible for most of the high correlations in the matrix. Clay minerals consist largely of aluminum oxides (Al_2O_3), and these particular clays are rich in iron (Fe) and magnesium (Mg). Metal ions such as zinc, chromium and nickel (Cr, Zn and Ni) easily adsorb to clays. We therefore hypothesize, based on the correlation matrix, that clay percentage was the most important *factor* that determines the chemical variation in the sediment samples. However, perhaps more factors are present. For instance, organic matter sedimentation (peat) might be an important factor since the borehole also contained organic sediments. Factor analysis is a family of techniques that allows us to find these factors and isolate them from the data.

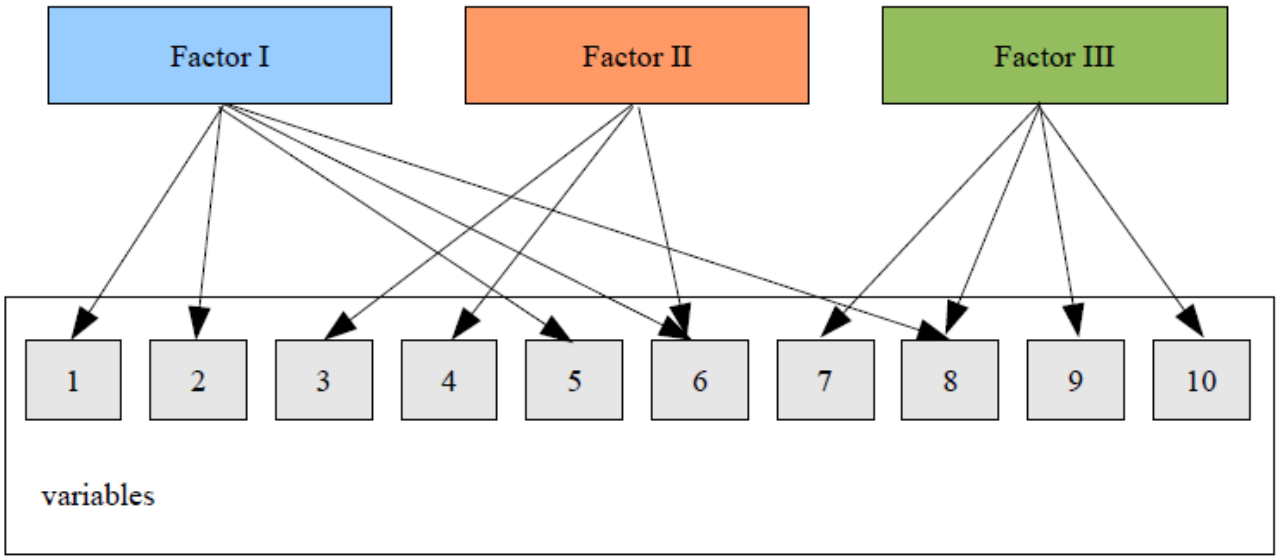

Figure 35: Schematical representation of factors, determining the variation of individual variables in a multivariate dataset. The arrows indicate which variable is influenced by which factor.

The idea behind factor analysis can be summarized as in Figure 35. In this example we assume a dataset with ten variables and three factors. The total variance in the dataset is explained by the 10 variables, but we can often find a smaller number of factors more efficiently explains a large part of the variance. When one factor has a strong influence on several variables (e.g. factor III and variables 7, 8, 9 and 10 in Figure 35) you can expect strong correlations between these variables. It is also possible that one variable is influenced by two or more factors (e.g. variable number 6, which in influenced by both factor I and II).

In the data in the table above, much of the variation in the sediment chemistry is determined by the clay content. This in turn is determined by the sedimentary environment: more clay is deposited at standing or slowly flowing water on a river floodplain than in parts with more rapid flow. Thus, an underlying factor could be the flow speed at which the sediment is deposited. However, since we do not have data on the velocity of the water in which the sediment is deposited, we cannot know this

for certain just by doing data analysis on this dataset, but we can interpret the meaning of a factor using reasoning like we did above.

The goal of factor analysis is to find factors that explain a part of the data and to find out which variables are influenced by which factors. In addition, we would like to know how *strong* the relationships between the factors and the variables are. This is expressed in Figure 36. Here, the thickness of the arrows indicates the strength of the relation. For example, variable 8 is very strongly determined by factors I and III. In factor analysis terms, this strength is called the **loading** of a factor on a variable.

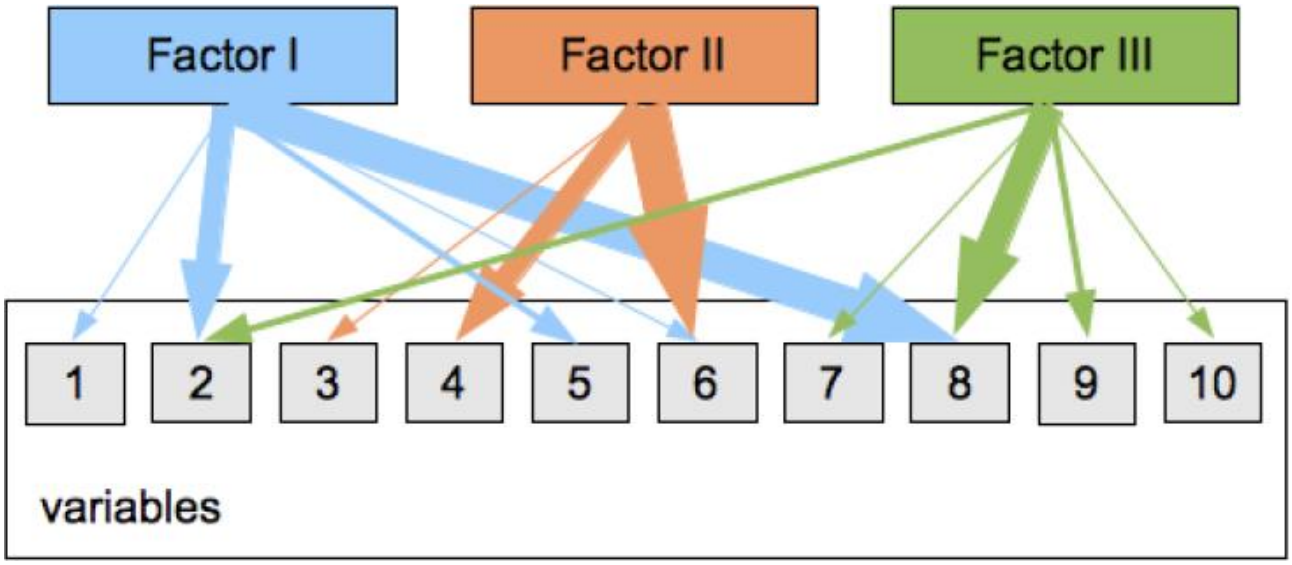

Figure 36: Factor loadings a measure of the strength of the relation between the factors and the variables. In this schematic, the loadings are indicated with the thickness of the arrows.

In the example above, factor I (e.g. flow speed of the water) has a very strong relation to the clay content. A stronger water velocity means less clay in the sample and therefore also less aluminum oxide, which is the main constituent of clay. At the same time, it negatively influences the sand content and the variables that are related to sand, such as silicon oxide (SiO_2). A 'loading' is simply a number that expresses the strength of this relation; the number can be both positive and negative. You can think of the loading as a correlation coefficient between a factor and a variable.

In the same way, you can relate the factors to the individual samples from which the chemical analyses originate (Figure 37). These relations can also expressed with positive and negative numbers, called **scores** in factor analysis. For instance in our example, if the flow speed factor has a strong positive score for a sample, this sample should originate from a sedimentary environment with weak currents, and should contain a lot of clay and little sand. To summarize:

Factor **loadings** relate the **variables** to the factors

Factor **scores** relate the **samples** to the factors.

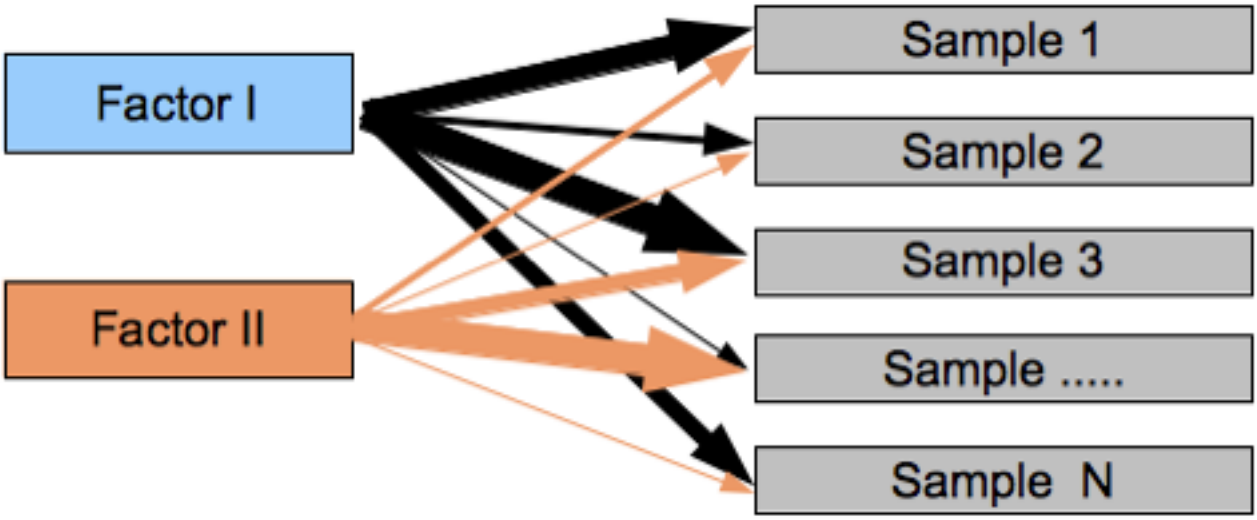

Figure 37: Factor scores: these show how each sample is influenced by the factors. Like in the previous figure, the thickness of the arrows indicate the strength of the relation.

The goal of factor analysis is to find the factor loadings and scores in the hope that you can determine which factors were at work to determine the variance of the data. This is a difficult question to answer, in particular when you do not have any idea what to look for, for instance in a large dataset and with poor theoretical knowledge. In the geochemistry example above, we could use geochemical background knowledge and our knowledge about the relationship between water velocity and sediment content to interpret the correlation matrix, but we are not always in such a luxury position. Note that the flow speed of the water was not even a variable in our dataset, we had to make it up by interpreting the structure of the data. Furthermore there is always random errors in the data which obscure the relations.

X.2 The lake of Whamsterdam: A numerical example of factor analysis

To illustrate how factor analysis works, let's look at another numerical example: The city of Whamsterdam is situated on a big lake. Unfortunately this lake is polluted by a number of industries. At a certain moment the brave citizens get sick and tired of all the dead fish floating around and they want their lake to be clean again. In the next round of elections they demand immediate action. When things are being discussed in the town council, the decision for action proves to be difficult: There are of course economic interests of the people earning their living in the factories. Which factory should be shut down first or forced to reduce its pollution? The council decides to do research on the pollution, mostly in an attempt to delay the decision (a common political strategy, unfortunately…). A consultancy company is hired to do this research. First, the environmental scientists of the consultancy ask the industries to tell which chemicals, and how much of each, they dump into the lake. Law-abiding as they are, the industries hand over the requested figures. Since the consultants are experienced environmental scientists, they don't trust the numbers and decide to go on a sampling campaign to get some independent data on the lake water composition. Boats set out on the lake to take a large number samples.

Below is the result of the questionnaire among the factories. There are three factories: I, II and III. They dump four different chemicals into the lake, labelled A, B, C and D.

- Factory I dumps 100 kg of chemical A, 0 kg of chemical B, 100 kg of chemical C and 50 kg of chemical D into the lake per day
- Factory II dumps 0 kg of chemical A, 100 kg of chemical B, 0 kg of chemical C and 100 kg of

chemical D into the lake per day.

- Factory III dumps 50 kg of chemical A, 0 kg of chemical B, 150 kg of chemical C and 0 kg of chemical D into the lake per day.

We can put this data in a matrix, with the rows denoting the chemicals dumped by each factory, and the columns the factories (Table 12)

*Table 12: Matrix **L** listing the amounts of chemicals (A, B, C and D) dumped into the lake of Whamsterdam by each of the three factories (I, II and III) in kg/day.*

	I	***II***	***III***
A	100	0	50
B	0	100	0
C	50	0	150
D	100	100	100

The scientists of the engineering company call this table the matrix of *loadings*, in short: matrix **L**. The sample data are also entered into a matrix. Here, the rows represent the measured amount of chemicals while the columns indicate the samples, numbered consecutively 1, 2, 3, etc (Table 13). Not all samples are shown. At least 250 have been taken and analyzed in the lab. The scientists call this table the data matrix, matrix **X**.

*Table 13: Matrix **X** listing the amount of chemicals A, B, C and D measured in samples 1, 2, 3, 4, etc. (not all samples are listed).*

	A	***B***	***C***	***D***
1	15	80	25	100
2	50	25	100	100
3	…	….	…	…
4	…	….	…	…

The contribution of every factory is not the same everywhere in the lake. The currents in the lake are not strong enough to mix the lake water completely. So, close to the waste outlet of factory I you would expect the contribution of factory II and III to be relatively small and that of Factory I to be the most important. Actually, the scientists can calculate the strength of this contribution from the distance of each sample location to the factory. This results in a matrix of weights, which gives the relative contribution of each factory to the lake water concentrations at each sample location. The scientists say that they have determined how much a factory *scores* on a certain sample location, and they call the result the matrix of factor(y) scores, matrix **S** (Table 14). Again, the rows represent the samples. The columns now represent the factories.

*Table 14: Matrix **S** showing the relative contributions (scores) of each factory (I, II and III) on the lake water*

composition at the locations of the samples (1, 2, 3, 4, …).

	I	***II***	***III***
1	0.1	0.8	0.1
2	0.25	0.25	0.5
3	…	….	…
4	…	….	…

The scientists now can calculate what the composition of each sample should be, based on the amount of chemical dumping stated by each factory and the relative importance of each factory's output for each sample. In this way they want to check whether the data on chemical effluents from the factories are right. The calculation can be made easily using the rows of matrix **S** and matrix **L**. Let's walk through this calculation step by step. Here is the calculation for sample 1:

We start with chemical A. The first row of **L** states that factory I dumps 100 kg, factory II 0 kg, factory III 50 kg. The contribution score of each factory to sample 1 is in the first row of matrix **S**. The contribution for factory I to chemical A at location 1 is therefore 0.1 x 100, that of factory II 0.8 x 0, factory III 0.1 x 50:

$$X_{1,A} = 0.1 * 100 + 0.8 * 0 + 0.1 * 50 = 15$$

For sample 1, chemical B we can use the same calculation, now taking the second row of **L** and the first row of **S**:

$$X_{1,B} = 0.1 * 0 + 0.8 * 100 + 0.1 * 0 = 80$$

For sample 1, chemical C and D:

$\boldsymbol{X}_{1,C} = 0.1 * 100 + 0.8 * 0 + 0.1 * 150 = 25$ (3d row **L**, 1st row **S**)

$\boldsymbol{X}_{1,D} = 0.1 * 100 + 0.8 * 100 + 0.1 * 100 = 100$ (4th row **L**, 1st row **S**)

Similar, for sample 2:

$\boldsymbol{X}_{2,A} = 0.25 * 100 + 0.25 * 0 + 0.5 * 50 = 50$ (1st row **L**, 2nd row **S**)

$\boldsymbol{X}_{2,B} = 0.25 * 0 + 0.25 * 100 + 0.5 * 0 = 25$ (2nd row **L**, 2nd row **S**)

$\boldsymbol{X}_{2,C} = 0.25 * 100 + 0.25 * 0 + 0.5 * 150 = 100$ (3d row **L**, 2nd row **S**)

$\boldsymbol{X}_{2,D} = 0.25 * 100 + 0.25 * 100 + 0.5 * 100 = 100$ (4th row **L**, 2nd row **S**)

This pattern of calculation is the same as multiplying matrix **S** with the transpose of matrix **L**: **L'**:

$$\boldsymbol{X} = \boldsymbol{S} * \boldsymbol{L}'$$

(72)

The scientists also should account for a matrix of random measurement errors, which occur in every laboratory determination of the chemicals:

$$\boldsymbol{X} = \boldsymbol{S} * \boldsymbol{L}' + \boldsymbol{e}$$

(73)

The error matrix **e** is the same size as matrix **X**, so every variable on every sample has its own error. Now, the scientists have a calculation of the samples and they can check whether the industries told the truth about their pollution.

This story can be translated into a factor model with the following elements:

- The factories are the factors
- The chemicals are the variables
- The samples taken by the consultancy scientists are the samples (rows in the data matrix)
- The matrix of loadings **L** says how the factors influence the variables
- The matrix of scores **S** says to what extent each observation is determined by a factor.

In real factor analysis problems, we only have matrix **X**, which lists the values measured for each variable in each sample. In factor analysis in particular, **L** is the thing we are looking for: the relation between factors and variables. **S** and **L** are missing, and have to be determined from **X**. In the equation above (equation 72), it looks like we have one equation with two unknowns (and then we are not even considering the errors **e** on the measurements), so you might think that it is impossible to solve this problem. Luckily, it is less hopeless than it sounds. Using some assumptions about the properties of the factors, we can find the factors from the data. Let's dive into the mathematics of how this works!

X.3 An outline of the mathematical basis and the terminology.

To find the factors, we first must realize ourselves what properties they should have:

1. It is usually assumed that factors act independently from each other. In Figure 36, factor I should not have any influence on factor II. In statistical terms, this means that they are *uncorrelated*. If the factors can be expressed in numerical values, they should have a correlation coefficient of zero (r = 0).
2. We must make assumptions about how they influence the variables. In factor analysis it is assumed that the relationship between a factor and a variable is very simple, comparable to a linear regression equation: We simply multiply the variables related to one factor by constants and add them up. This is also known as a *linear combination.*
3. The influence of the factors should be distinguishable from the random errors in the data.

Based on these assumptions, we can find a way to derive the factors from the original data. This seems somewhat awkward, and the procedure may be difficult to understand. It will be explained here largely using graphical examples without an in-depth mathematical treatment.

We start with the property of no correlation: If we remove correlation between variables from our dataset, we may end up with something like the factors we are looking for. A dataset without correlation should indicate more directly to what extent it has been influenced by a single factor. To understand how we remove correlation from the data, we need to go back to sections VI.3: Reduced Major Axis Regression (RMA) and IX.1 The multivariate normal distribution. This step is crucial for understanding what is happening mathematically in factor analysis.

In VI.3: Reduced Major Axis Regression (RMA), we have seen that regression lines between two variables can be drawn in several ways. If we assume no dependency between the variables, a regression line called the 'principal axis' is used (or its close relative which is more easily computed, the reduced major axis). In IX.1 The multivariate normal distribution, we learned that a

multivariate population can be described by the multivariate normal distribution. The probability contours of the surface of this distribution are elliptical. The largest (major, principal) axis of these ellipses is the same as the principal axis. Note: When we have more than two variables, this ellipse is not a simple ellipse in two dimensions with one long and one short axis, but an elliptical hypersurface with as many dimensions and as many axes as variables. In VI.3: Reduced Major Axis Regression (RMA), is also demonstrated that you can move (translate) and rotate the coordinate axes of the data plots in such a way that all correlation between the data is lost. This is done by moving the origin of the coordinate axes to the mean of the data points, and rotating the axes parallel to the principal axes. This is shown in Figure 20, and more extensively in Figure 38.

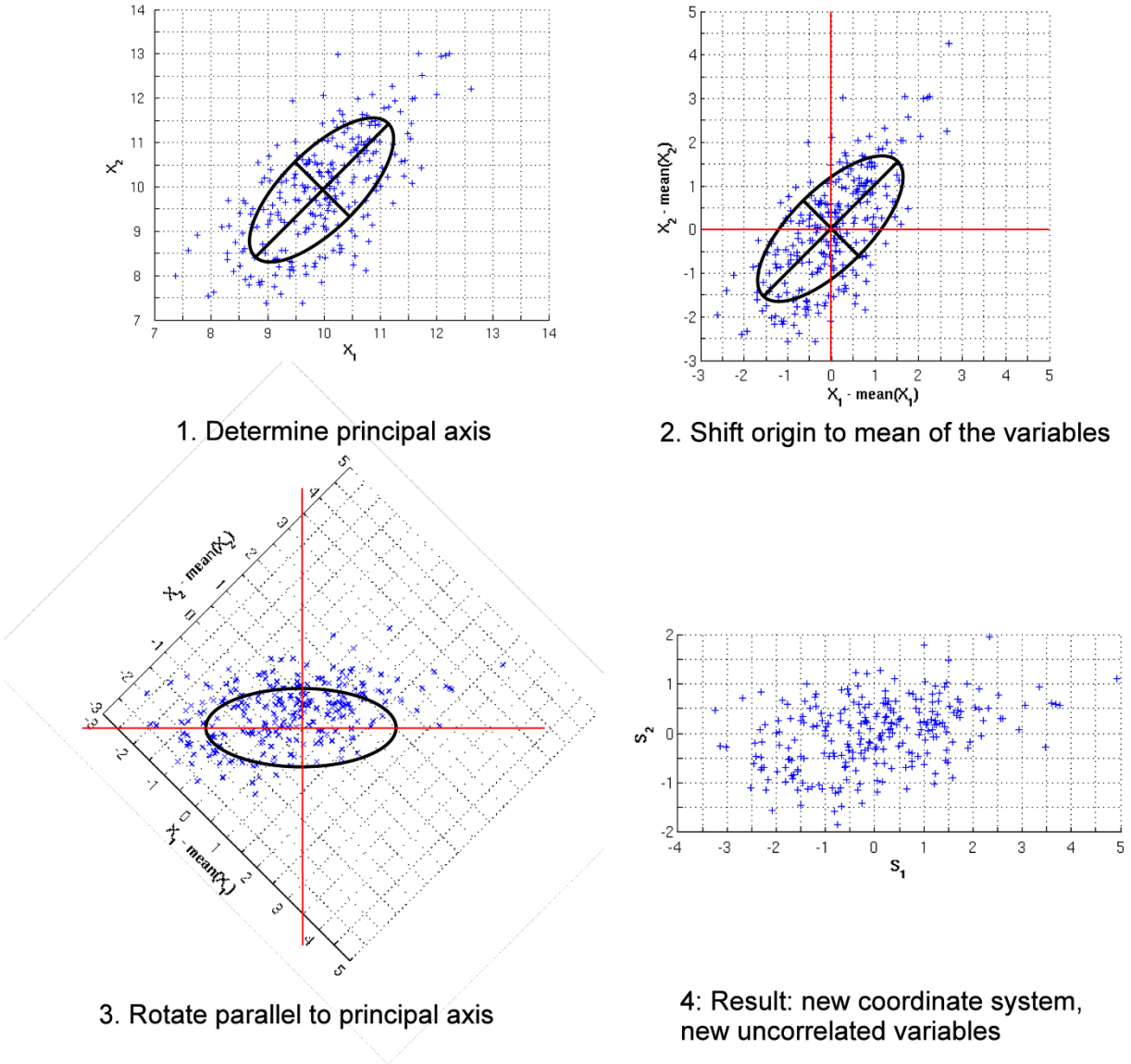

Figure 38: Change of the coordinate system of multivariate data to create uncorrelated variables from the original correlated data. The ellipse represents the 1 σ (standard deviation) horizontal cut through the multivariate normal distribution function.

In the top left of Figure 38 we see a scatterplot of the original data. Also an ellipse is drawn which represents a horizontal cut through the multivariate normal distribution for this population (see example of such a distribution in 3D in Figure 28), with its major (principal) axis, and minor axis perpendicular to it. In the top right of Figure 38, the origin of the scatterplot has been moved to the center of the data by subtracting the column mean (mean value for each variable) from every row

(for every sample) in the datamatrix. In the lower left, the coordinate system has been rotated to align with the principal axis of the data. This has been done by performing a second transformation to the data, which we will discuss below. On the lower right, we see the transformed data relative to the new axes. The center (mean) of all the data points now lies at the origin. The horizontal axis now contains the largest variation and the vertical axis the smallest variation of the dataset. This can be seen easily from the spread of the datapoints along the axes. If you would calculate a correlation coefficient from these transformed data, it would be zero.

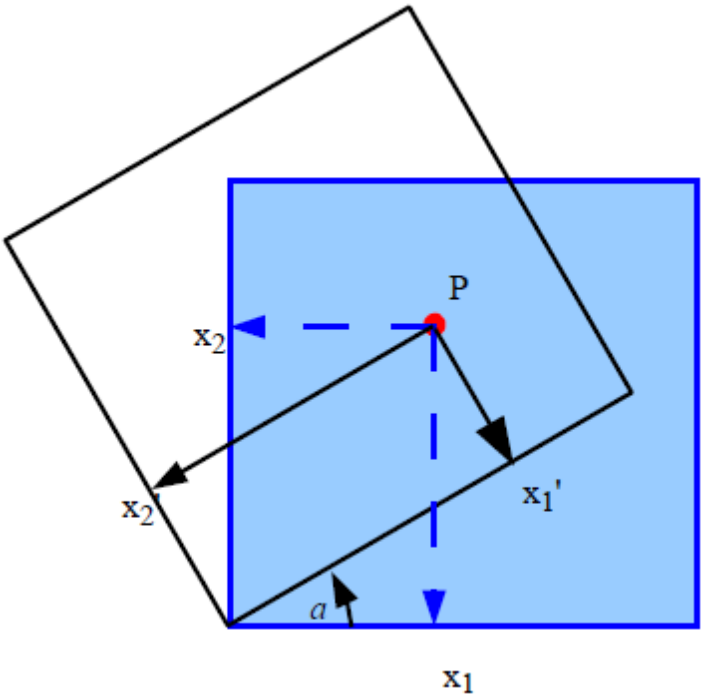

Figure 39: Rotation of coordinate axis and transformation of old coordinates of point P (x1,x2) to new coordinates in the rotated coordinate system P(x1',x2')

Now the question is how to rotate the coordinate system of the data. If we have a point P plotted in a Cartesian coordinate system, we can rotate the axes of the coordinate system. Point P then will have a new set of coordinates relative the new coordinate system, which can be derived from the old coordinates and the rotation angle (see Figure 39). We again use a two-dimensional example here to keep it simple. If the rotation angle is *a* and the old coordinates are x_1 and x_2, then the new coordinates x_1' and x_2' can be derived by the following set of equations:

$$x_1' = x_1 \cos a + x_2 \sin a$$
$$x_2' = -x_1 \sin a + x_2 \cos a$$

(74)

This operation can in fact be expressed as a matrix multiplication:

$$\begin{bmatrix} x_1' \\ x_2' \end{bmatrix} = \begin{bmatrix} \cos a & \sin a \\ -\sin a & \cos a \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

(75)

The matrix with sines and cosines is called a rotation matrix. Of course, such a rotation matrix can be easily expanded to more than two dimensions. We can also see another feature of such a rotation: the new coordinates are a linear combination of the old ones. So, when we perform the transformation of the coordinate system shown in Figure 38, we get new variables that have properties which factors we are looking for should have: They are uncorrelated and they can be

produced from a linear combination of the original variables.

The last question we should answer is: How do we find the rotation matrix which aligns our data coordinate axis with the principal axis? In other words (and in the case of 2 dimensions): What is the angle *a* that we need to align the coordinate system with the principal axis? To calculate this, we need to perform a matrix manipulation that will yield the **eigenvectors** and **eigenvalues** of the data. We will not deal with the mathematics of how to do this, but it is treated more in-depth by Davis (2002[1]) and in this blog post in the section on matrix algebra and in VII.7 Extra reading: Eigenvectors and Eigenvalues below.

For now, it suffices to say that every square matrix (such as the correlation matrix of our dataset, see e.g. Table 11) can be associated with a matrix of eigenvectors and corresponding eigenvalues. If we calculate the matrix of eigenvectors from the covariance matrix of the data, the result is the rotation matrix that we need to align the coordinates of our data to the principal axes. Moreover, the associated eigenvalues are the variances of the transformed data along the new coordinate system. In the case of a two-variable data set, it is the maximal variance along the principal axis and the minimal variance along the minor axis perpendicular to it. The eigenvalues indicate how long this axis is. In other words: They indicate how much variation is in the dataset in the direction of the axis (or "factor"). In multivariate factor analysis, the eigenvectors and eigenvalues are ordered with decreasing variance: The first axis always contains the highest amount of variance, the second contains the second highest amount, etc. To summarize:

- The **eigenvectors** are the *rotation matrix* that rotates the coordinate system of the data to align it with the direction of maximal variance: the principal axes of the data.

- The **eigenvalues** indicate the *magnitude of the variances* along each of the principal axes.

- We use the **eigenvectors** to determine the *direction* of the principal axes and we then use the **eigenvalues** to estimate the *relative importance* of each principal axes.

X.4 Principal component analysis.

Principal component analysis (PCA hereafter) is the simplest form of factor analysis. In fact, it is usually not even considered as a type of factor analysis, but more an exploratory tool to find out how much factors there could be. It is not more than a rotation of the coordinate system of the data space to align it with the principal axes of the data population. We will discuss principal component analysis using an example, but first we highlight the steps we need to take to do a principle component analysis below:

1. *The first step* is to standardize the data matrix. Standardization already has been discussed in section VIII.3 The correlation matrix. From each column entry, the mean of the column is subtracted and then the entry is divided by the column standard deviation ($z_{ij} = \frac{(x_{ij} - \mu_j)}{\sigma_j}$). This puts all the variables on the same scale and removes the effects of different measurement scales. For PCA, this has the additional advantage that a part of the transformation of the coordinate system of the data space already has been done: the origin is moved to the centre of the data.
2. *The second step* is the calculation of the correlation matrix, as described in VIII.3 The correlation matrix.
3. *The third step* is to calculate eigenvalues and eigenvectors of the correlation matrix (see VII.7

[1] Davis and Sampson, *Statistics and Data Analysis in Geology*.

Extra reading: Eigenvectors and Eigenvalues). In the terminology of the PCA method, the eigenvectors are known as the **principal components**. The columns of the eigenvector matrix, which contain the parameters relating all the principle components to the original variables in the dataset, are known as the principal component **loadings**. Below, we will discuss the results in the example, and show what you can read from these results.

4. *The fourth step* is to calculate the principal component **scores**. These are the transformed data, or in other words: The coordinates of the original (standardized) data with respect to the new, rotated coordinate system. The principle component scores therefore relate each sample in the dataset to each principle component. Again, their interpretation will be discussed below.

Now for the example: This is the same dataset of river deposits used in the example of X.1 The theory behind factor analysis: Loadings and Scores, although it has been simplified by deleting a few variables (the rarer metals: Zn, Cr etc.). Two other variables have been added: The sand and silt content. We assume that the chemistry of the sediment may tell us something about the different sources of the river sediments we find in the sequence of samples we have took from the core. In factor analysis terminology: the assumption is that the variability in chemical composition of the sediment is determined by factors that represent the sediment sources. The general characteristics of the sediment sources in the area are well known. The following sources may be expected:

1. *Tertiary marine clays and sands*. The clays contain high amounts of the Fe- and Mg-rich clay mineral group of the smectites. In addition, the mineral glauconite is often abundantly present. This is also a clay mineral which contains Fe and K. It gives the sediments a characteristic greenish color. Upon first inspection of the core some of the units appeared more greenish than others, hinting that this component might be important.
2. *Glacial and fluvioglacial sediments*. These originate from the Scandinavian ice sheet which invaded the area during the Saalian glaciation (238.000 to 126.000 years ago). These are generally sediments with high amounts of calcium carbonate and fresh, unweathered igneous rocks with feldspars from the Scandinavian shield. Characteristic elements in feldspars are Na, K, and Ca. The carbonate obviously contains high amounts of Ca. The grainsize range is very large, from boulders to clay, but silt dominates.
3. *Loess*. This has likely been deposited during the last glaciation (Weichselian; 115,000 to 11,700 years ago) when the area was unglaciated but had a periglacial (tundra-like) climate subjected to dust storms from the ice margin. Chemically, it is similar to the glacial sediments, but it consists mainly of silts.
4. *Middle Pleistocene fluvial sediments*. These consist mainly of quartz-rich, weathered sediments ranging from sand to gravel.

This description shows that the problem is by no means simple. There are no factors or sediment sources that can be detected by one single variable in the data set. Several factors influence the same variable. For instance, the content of the element K may be enhanced by sediment source 1, 2 and 3. Moreover, our list of factors may not be complete. The sediment chemistry could be influenced by other processes besides the sediment sources such as weathering in floodplain soils after deposition and other chemical alteration processes such as the formation of chemical deposits by circulating groundwater.

The correlation matrix for this dataset is given in Table 15. Knowing the number of samples (n = 55), we can calculate which correlation coefficients are significant with a high significance level (p = 0.01)

using formula 63 in VIII.3 The correlation matrix. The have been displayed with a light red background in the table. The pattern of significant correlations between clay, Al_2O_3, Fe_2O_3 and MgO are striking, pointing to the presence of smectite clays. However, the remaining pattern is more difficult to interpret from the correlation matrix alone.

Table 15: Correlation matrix of dataset consisting of 53 sediment samples. Correlation values indicated in light red are statistically significant at the 99% confidence level ($p < 0.01$).

	Al_2O_3	Fe_2O_3	MgO	CaO	Na_2O	K_2O	$CaCO_3$	Clay	Silt	Sand
Al_2O_3	1	0.7835	0.8704	-0.0924	0.0091	0.6420	-0.1875	0.8256	0.1411	-0.5988
Fe_2O_3		1	0.7027	-0.0058	-0.2056	0.3041	-0.2535	0.7959	0.2492	-0.6373
MgO			1	0.0801	0.3173	0.7628	0.0108	0.6073	0.3546	-0.6454
CaO				1	0.0930	-0.1624	0.7693	-0.0625	0.2584	-0.2449
Na_2O					1	0.6638	0.4364	-0.4132	0.4250	-0.0971
K_2O						1	0.1524	0.2269	0.2748	-0.3507
$CaCO_3$							1	-0.3342	0.1510	0.0217
Clay								1	0.0125	-0.5859
Silt									1	-0.7441
Sand										1

After calculating the eigenvalues and eigenvectors of the correlation matrix we obtain 10 eigenvalues and 10 eigenvectors, the same amount as the variables. However, from the results we can determine whether we could do with less than 10 principal components. As noted in the previous section, the eigenvalues represent the amount of variance of the data along the principal axes (= components) of the data. We can use this knowledge to see how much of the total variance in the data is represented by each principal component.

Table 16 lists the eigenvalues. In the first column, the principal components are given a number in order of importance. In the second column the eigenvalues are given. The total variance of the data is simply the sum of the eigenvalues. In this case equal to 10. Since we have standardized the data, the variances of the original variables all have reduced to 1, resulting in a total variance of 10 for the entire dataset. The next column expresses each eigenvalue as a percentage of their total, and the last column gives the cumulative percentages.

Table 16: Table showing the eigenvalues of all 10 principal components in the river sediment dataset.

Principal component	eigenvalue	percentage of total	cumulative percentage of total
1	4.3550	43.5%	43.5%
2	2.5004	25.0%	68.6%
3	1.5231	15.2%	83.8%
4	1.0229	10.2%	94.0%

Principal component	*eigenvalue*	*percentage of total*	*cumulative percentage of total*
5	0.2283	2.28%	96.3%
6	0.1532	1.53%	97.8%
7	0.0898	0.90%	98.7%
8	0.0538	0.54%	99.3%
9	0.0468	0.47%	99.7%
10	0.0267	0.27%	100%

From Table 16 we can see that the first principal component contains most of the variance: 43.5%. The amount of variance contained in the second (25%), the third (15%) and the following principal components decreases with each step. Together, the first three principal components make up 83.8% of the total variance. Adding the fourth principal component, we have already captured 94% of all the variance of the data. The remaining principal components add only small percentages to this.

The fact that only four principal components describe nearly all variance in the data suggests that we can explain much of what is going on the dataset with only four principal components, or factors. Maybe we can even make do with less, for example three, components. This is considerably less than the 10 variables we started with. So, for any further data processing these three or four principal components may be used instead of the original variables. The remaining principal components probably represent only uncertainty in the data.

Now you might think: How do I choose how many principal components I should keep? This again one of these questions that pure data analysis cannot help you with, unfortunately. To determine how many principal components you want to retain in a dataset, you must ask yourself two questions:

1. Do I have any pre-knowledge about how many factors I expect to be important in my dataset?
2. How much of the variance in my dataset do I expect to leave "unexplained" while still feeling confident that I have gained all the information I need from my dataset?

In the example dataset, we had a preconceived idea that there would be four sources of sediment in our samples. Therefore, retaining 4 principle components seems to make sense. Also, given the type of measurements that went into this dataset (grain size analysis and chemical analyses), we are probably not surprised that 7% of the variance in our dataset is left unexplained. 7% seems like a realistic estimate of the relative uncertainty we might expect on these types of measurements. Finally, we can see from Table 16 that the fifth principal component explains significantly less variance (2.3%) than the fourth (10.2%). Such a "drop" in the amount of variance explained moving from one principal component to the next is a good sign that we have "exhausted" our dataset and that the remaining components are not very important. All these considerations may help you to determine which principal components you want to retain in your dataset, but this remains a scientific decision. Once again, data analysis is not going to do the research for you!

Table 17: Table containing the eigenvectors, or the loadings of all 10 principal components on all 10 original

variables in the river sediment dataset.

	principal components									
Variable	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
Al_2O_3	0.4431	-0.1156	0.0796	0.2448	-0.1573	-0.0556	-0.1597	0.4697	-0.6439	0.1884
Fe_2O_3	0.4036	-0.1861	-0.1695	0.0230	0.7839	0.3200	0.1619	-0.1135	-0.0747	-0.1076
MgO	0.4427	0.1006	0.1136	0.1744	0.1561	-0.5214	-0.3193	-0.1998	0.4163	0.3729
CaO	0.0088	0.3694	-0.6017	0.2645	0.0451	-0.4268	0.0336	0.0618	-0.1117	-0.4794
Na_2O	0.0660	0.5131	0.4185	-0.0547	0.1476	-0.1791	0.6765	0.1776	-0.0466	0.0956
K_2O	0.3156	0.2425	0.4770	0.1907	-0.1766	0.2628	-0.2391	-0.2221	0.0452	-0.6074
$CaCO_3$	-0.0588	0.5115	-0.2474	0.4250	-0.0569	0.5501	-0.0851	-0.0122	0.0980	0.4126
Clay	0.3727	-0.3111	-0.1925	0.1682	-0.3726	0.1289	0.4189	0.2998	0.5177	-0.1009
Silt	0.2155	0.3371	-0.1398	-0.6694	0.0449	0.1376	-0.3173	0.4596	0.1907	-0.0559
Sand	-0.3897	-0.1066	0.2636	0.3779	0.3705	-0.0349	-0.2168	0.5837	0.2753	-0.1499

Of course, we need to understand what these principal components might represent. To do so, we must do some interpretative work by looking at the eigenvectors. These are given in Table 17. Each column represents an eigenvector belonging to a principal component. The values in the columns are also called '*principal component loadings*'.

To explain what these numbers in the eigenvectors tell us, remember that the eigenvector matrix is used to transform the data from their original variable coordinate system to a new principle component coordinate system, using the matrix multiplication of equation 75 in X.3 An outline of the mathematical basis and the terminology.. The transformed data points are known as '*principal component scores'.*

We have explained what these terms mean above, but writing out the calculations on how we might use the numbers in the eigenvectors / principal component loadings can help to understand how to interpret these results. If we call the matrix of principal component loadings / eigenvectors (Table 17) **L**, the original data (after standardization) **Z**, and the matrix of principal component scores **S**, then the following function shows how we can obtain the principle component scores from the original data and the principle component loadings (or eigenvectors):

$$\boldsymbol{S} = \boldsymbol{Z} * \boldsymbol{L}$$

(76)

Notice that this is basically the same function as equation 75 but expressed in more general terms (and for larger matrices instead of just two points). Writing out the matrix multiplication, the principal component score for sample *i* on principal component *j* for a dataset with *m* variables can be calculated as:

$$\boldsymbol{S_{ij}} = \boldsymbol{L_{1j}Z_{i1}} + \boldsymbol{L_{2j}Z_{i2}} + \boldsymbol{L_{3j}Z_{i3}} + \ldots + \boldsymbol{L_{1m}Z_{im}}$$

(77)

Below we see an example for sample number 3 of the data set above. The original, unstandardized data for sample 3 are:

$$\boldsymbol{X_3} = [5.70 \quad 1.12 \quad 0.61 \quad 2.21 \quad 1.02 \quad 1.55 \quad 3.50 \quad 3.17 \quad 21.50 \quad 26.08] \tag{78}$$

Here, the values are, from left to right, the measurement results for Al_2O_3, Fe_2O_3, MgO, CaO, Na_2O, K_2O, $CaCO_3$, Clay content, Silt content and Sand content, respectively. All values are weight percentages. The standardized data (z-scores) are:

$$\boldsymbol{Z_3} = [-0.54 \quad -0.97 \quad -0.17 \quad 0.56 \quad 0.94 \quad 0.20 \quad 0.97 \quad -0.72 \quad -0.53 \quad 0.36] \tag{79}$$

To obtain the score for sample 3 on principal component 1, every standardized variable is multiplied with the corresponding loading value from the 1st column of **L** (in red) and summed:

$$\begin{aligned} \boldsymbol{S_{3,1}} = {} & 0.44 * -0.54 + 0.40 * -0.97 + 0.44 * -0.17 + 0.01 * 0.56 + 0.07 * 0.94 + 0.32 \\ & * 0.20 + -0.06 * 0.97 + 0.37 * -0.72 + 0.22 * -0.53 + -0.39 * 0.36 \\ = {} & \boldsymbol{-1.16} \end{aligned} \tag{80}$$

In a similar way, the score on principal component 2, 3, etc. are obtained by taking the values from the 2nd, 3rd, etc. column of **L**. Finally, the scores for sample 3 on all principal components are:

$$\boldsymbol{S_3} = [-1.16 \quad 1.47 \quad 0.33 \quad 0.73 \quad -0.25 \quad -0.19 \quad 0.30 \quad -0.22 \quad -0.03 \quad 0.08] \tag{81}$$

As you have seen above, we call the values in **L** *loadings*. The numbers in the eigenvector (columns of **L**), or '*loadings*' can be seen as *weights*, that determine how important each original variable is for a particular principal component. For example, if the variable $CaCO_3$ has a very high number in the eigenvector of principal component 2, it means that the scores of all samples with high amounts of $CaCO_3$ will tend to be high on principal component 2. A high positive or strongly negative loading, means the corresponding variable is very important in determining the principal component. One can say that ***the loadings relate the variables to the principal components.*** For ease of interpretation, the principal component loadings are often displayed graphically in bar graphs. This gives a quick overview of the differences in weights of the variables, as in Figure 40, Figure 41 and Figure 42. Now, let's interpret our principle components:

The first principal component (PC) has high weights/loadings on Al_2O_3, Fe_2O_3, MgO, K_2O and clay, and a very negative weight on sands (Figure 40). The weights on clay and sand tell us that samples that score high on PC 1 will be clay-rich samples, which is also confirmed by the high weight on Al_2O_3. The weights on Fe_2O_3, MgO and K_2O suggest that these clays are smectite clays and could also contain glauconite (See the description of the possible sediment sources at the start of our example). Thus, it is likely that the first principal component says how much of the sediment is derived from the Tertiary clays.

The second principal component has high loadings on CaO, $CaCO_3$, Na_2O and silt. The glacial sediments and loess contain much silt and calcium carbonate, so it is likely representing these sediment sources. The Na_2O indicates feldspars, which are common in unweathered igneous rocks. Given the high loading on the silt fraction, it seems probable that PC2 represents the loess component in our sediments.

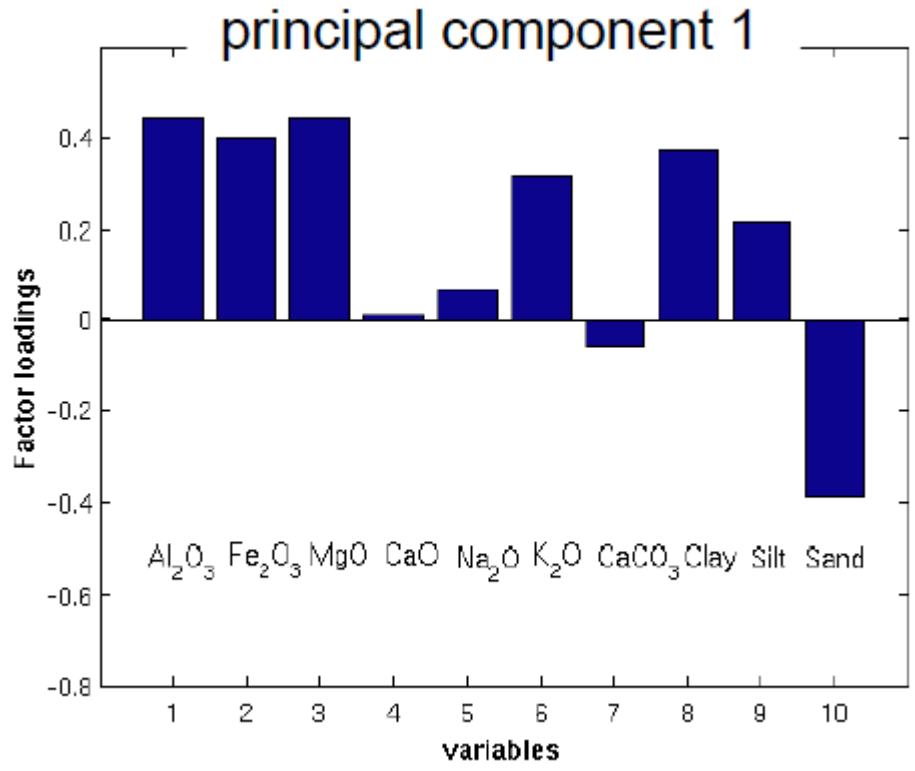

Figure 40: Principal component loadings of the first principal component of the example data set displayed as bar graphs

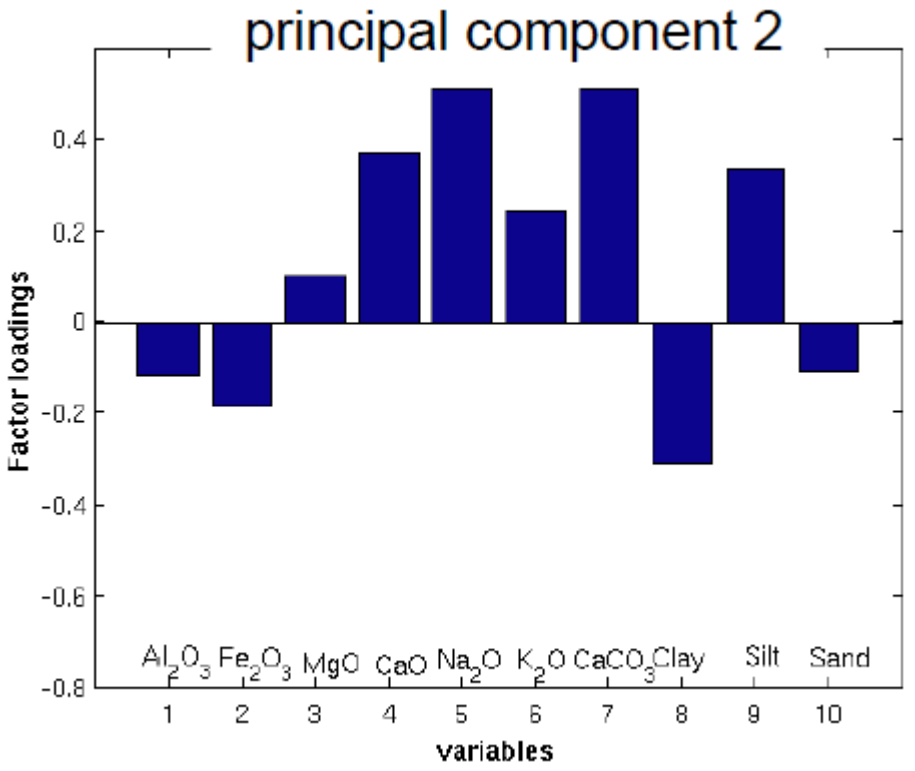

Figure 41: Principal component loadings of the second principal component of the example data set displayed as bar graphs

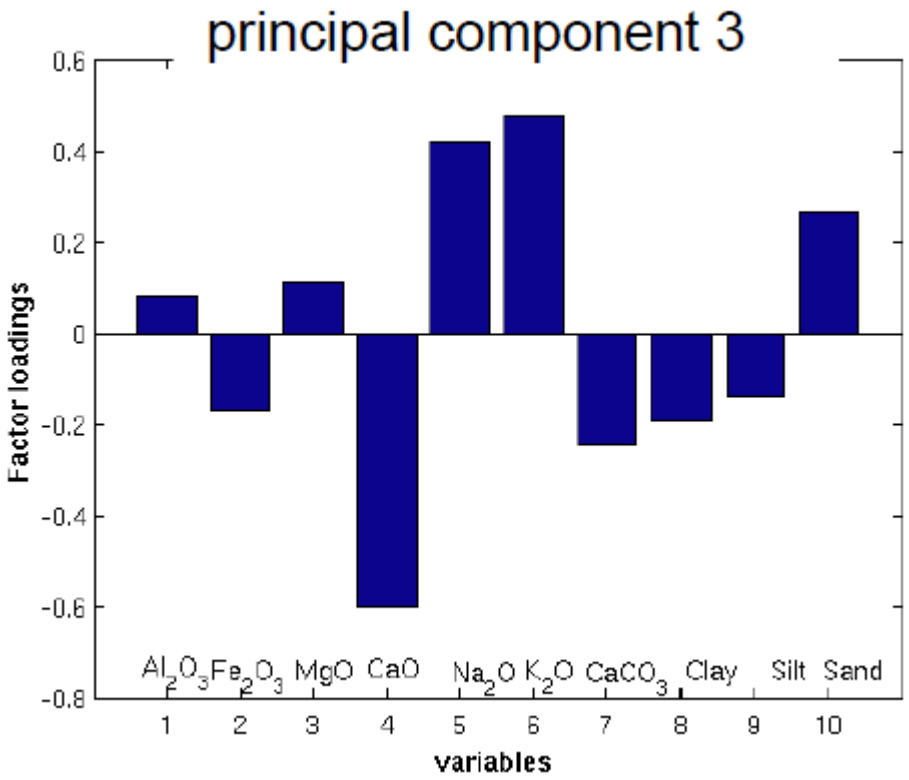

Figure 42: Principal component loadings of the third principal component of the example data set displayed as bar graphs

The third principal component is more difficult to interpret. It has a positive weight on sand and K_2O. This may represent glauconite-rich Tertairy sands, although this does not explain the strong negative weight on CaO and positive weight on Na_2O.

In general, the higher principal components are difficult to interpret. They contain less and less variance of the dataset, and therefore are much more likely to represent only meaningless random errors or a mix of several minor components. An interpretation of the first principal components may also be difficult. In this example, we could rely on some geochemical and geological background knowledge. Without this knowledge it would have been difficult to give any meaning to the principal components. In a case like this, the principal component analysis might help in drawing up hypotheses about the processes behind the data. We can then gather more data or carry out additional statistical analyses to test these hypotheses.

Now, let's look at the *principal component scores*. Remember that the principal components represent new axes, made by combining the original variables, on which you can plot the observations. Figure 43 shows a scatter plot of the scores of all samples on the first two principal components. This plot is obtained by plotting the values in the first two columns of matrix **S** against each other. Notice that these data look very poorly correlated. While, in the original data, high correlation coefficients occur between variables, by definition, a correlation matrix of **S** would contain only 0's.

However, by looking at the loadings, we now also know the meaning of the principal component axes. They represent processes that determine the composition of each sediment sample. The first principal component (horizontal axis in Figure 43) represents the admixture of Tertiary clays, and the second principal component (vertical axis) the amount of glacial material added to the sediment. Knowing this, we can interpret the origin of the samples. For instance, the two points in the lower right corner should be samples that largely consist of clay, and the topmost points represent samples

that largely consist of silt derived from glacial sources or loess. The values in the first column of S indicate the 'score' of any sample on the first principal component, or the 'Tertiary clay' axis. ***The principal component scores relate the observations to the principal components.***

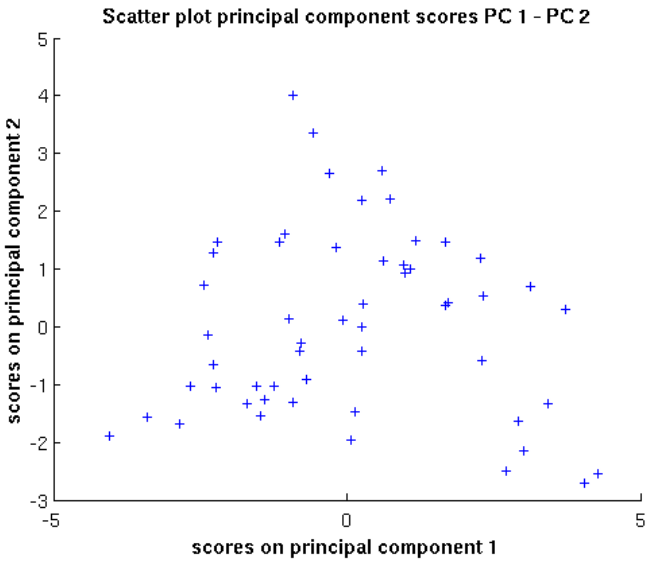

Figure 43: Scatter plot of principal component scores.

With our principal component analysis we now have:

1. Reduced the large number of original variables to a smaller number of factors. We might be able also to reduce the number of variables to be measured on further sediment samples, which saves time and money.
2. A better understanding on the probable causes of the variation in the data.

Principal component analysis is often a preparation of further data analysis. True factor analysis could be the next step. In true factor analysis, a decision is made on the number of factors that should be present. In our example, we might decide on three or four factors. The rest is considered as uncertainty (or “noise”) in the data. Next, a mathematical manipulation is used to optimize the loadings matrix. Usually, the pattern of loadings then becomes considerably clearer. Another step may be classification of the samples into groups based on the new factors. You have already seen some techniques for doing this in IX. MULTIVARIATE DISTRIBUTIONS AND CLASSIFICATION.

With the principal component analysis, the dimensions of the dataset can be reduced considerably, because we can continue our data analysis project using the principal component scores matrix (e,g, using only the columns of matrix **S** that represent the first three or four principal components) instead of the (bigger) original data. Many more complex data analysis operations are *computationally expensive*, meaning they take a lot of memory or computer processing to work. For the example we have discussed here, with 55 samples of 10 variables, the *data reduction* from 10 variables to 3 or 4 components may not seen very significant, but for large datasets (think of the data used by companies like Google and Amazon), this type of reduction can save a lot of time and money. It reduces the number of datacenters needed for data processing too, so in a way it is also good for the environment!

X.7 Take Home Messages

- **Factor analysis** is a family of data analysis operations that aim to reduce multivariate datasets with many correlated variables to datasets with a smaller number of uncorrelated **factors**.
- Factor analysis can also find structures in multivariate data that helps us to interpret what information the data contains.
- **Principal Component Analysis** is the most common and least complex form of factor analysis. Its aim is to transform the data to align it around principal axes that capture the highest amount of variance in the least number of components. This transformation is done using the **eigenvectors** of the correlation matrix of the dataset.
- The principal component **loadings** relate the **variables** to the principal **components**.
- The principal component **scores** relate the **observations** to the principal **components**.

VII.7 Extra reading: Eigenvectors and Eigenvalues

To find out how to rotate the data to align the axes of the dataset to the highest variance, we need to calculate the **eigenvectors** from the covariance matrix of the data. We have encountered the correlation matrix in VIII.3 The correlation matrix, but as a reminder, this matrix contains the correlation coefficients between each combination of variables in the dataset. So, for a dataset with *n* samples of *p* variables, the correlation matrix looks like:

$$\begin{vmatrix} cov(p_1,p_1) & cov(p_1,p_2) & \dots & cov(p_1,p_p) \\ cov(p_2,p_1) & cov(p_2,p_2) & \dots & cov(p_1,p_p) \\ \dots & \dots & \dots & \dots \\ cov(p_p,p_2) & cov(p_p,p_2) & \dots & cov(p_p,p_p) \end{vmatrix}$$

XI. TIME SERIES

XI.1 Basics

A time series is a series of observation data ordered in time. In time series analysis, time is almost always the independent variable, and the observation data are dependent variables. The time axis is usually displayed as the horizontal axis when graphing a time series. An exception are geological time series derived from vertical sections such as data collected at an outcrop or from a drill core. These are very often displayed vertically, with the time or depth axis as the vertical axis. The observations in a time series may have been taken at regular intervals, as is often the case with meteorological, hydrological, or economic data (see Figure 44).

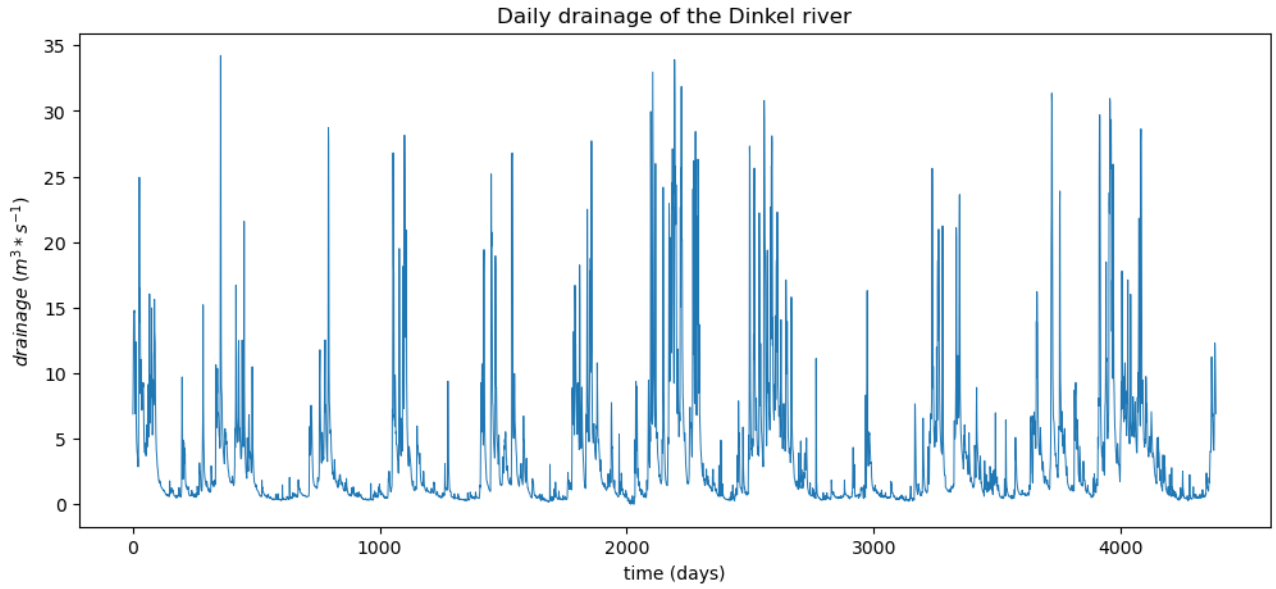

Figure 44: Discharge time series with time on the horizontal axis and equal-sized time steps of one day throughout.

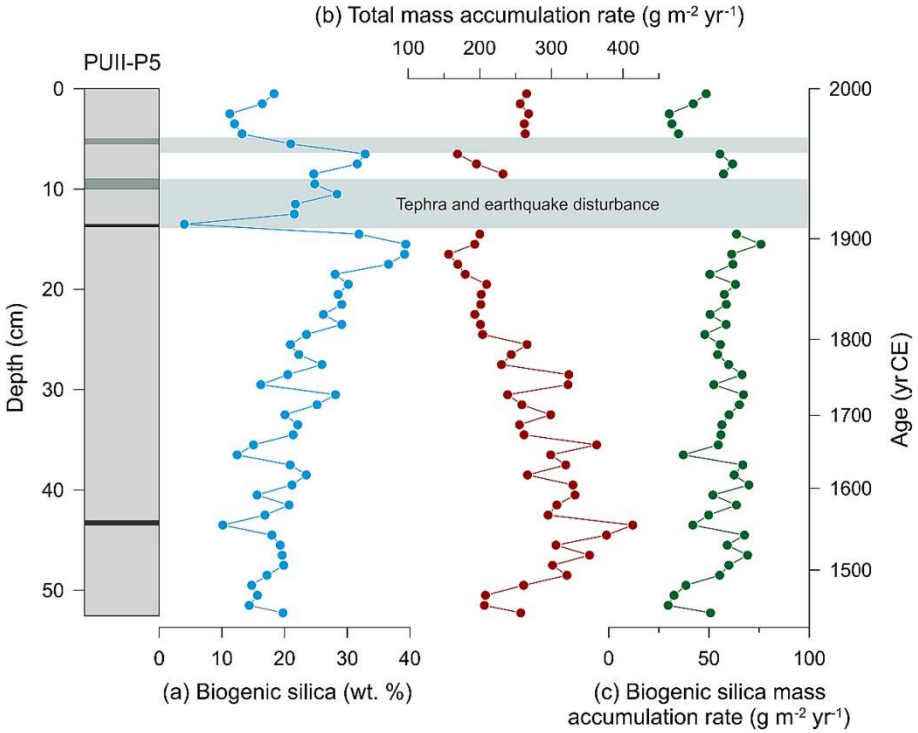

Figure 45: Typical geological time series: the biogenic silica concentrations and deposition rates of sediment in a

vertical section through lake sediments. In geological sections the vertical axis is often de depth axis. Here the time steps are unequal (notice how the steps between the age markers are different). Source: Bertrand et al., 2024[1]

In case of regular observation intervals, we only need a vector containing the observed values, the start time, and the time step to describe the time series completely. However, very often in the Earth sciences a regular interval is impossible. In geological sections or drill cores the depth axis is usually converted into a time axis by interpolating between levels of known age derived from dating methods. Since sedimentation rates are almost never constant, this inevitably leads to irregular time steps between the observations (see Figure 45). Such time series consist of two data vectors: one with the observations, and one with the corresponding time of observation according to the age model.

Unfortunately, most time series methods in statistics assume equal time steps. To convert our time series into a time series with an equal time step, we need to apply **interpolation**. There are several interpolation methods that can be used to obtain a time series with regular steps, but the most common, and simplest, one is **linear interpolation**. In linear interpolation we calculate the value from an unknown point from the nearest two points on both sides simply by drawing a straight line between the two known points (Figure 46).

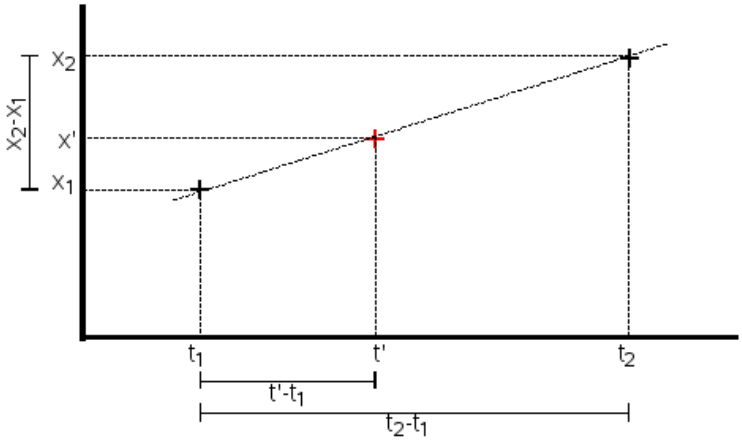

Figure 46: Linear interpolation. The red point is the point for which we know the t value (t`) but do not yet know the value. Its value (x`) is interpolated from its two nearest neighbors at t_1 and t_2 and their values (x_1 and x_2)

If we want to find the value of a point P' that is situated at time t' between two points (P_1 and P_2) of which we know the time values (t_1 and t_2) and the values for the dependent variable (x_1 and x_2), we can calculate the unknown value x' using the formula:

$$x' = \frac{(x_2 - x_1)(t' - t_1)}{t_2 - t_1} + x_1 \qquad (82)$$

Here t_1 and t_2 are the times of the two neighbouring known observations, x_1 and x_2 are their observed values, t' is the point on the time axis where a value hat to be interpolated, and x' is the interpolated value.

We call linear interpolation the "simplest" method because it requires the smallest number of

[1] Sebastien Bertrand et al., "Inorganic Geochemistry of Lake Sediments: A Review of Analytical Techniques and Guidelines for Data Interpretation," *Earth-Science Reviews* 249 (February 1, 2024): 104639, https://doi.org/10.1016/j.earscirev.2023.104639.

assumptions about the shape of the data between the two points we already know. Remember from IV.2 How to get a curved regression line - higher order polynomials. that a curved line requires more parameters to constrain than a straight line. Since we generally don't know what happens between the two points we use for our interpolation, the safest choice is to assume that the relationship between the dependent variable and time is linear. If we make things more complicated, we run the risk that we are *overfitting* with respect to the amount of information we have. Other methods of interpolation exist, such as spline interpolation that calculates the unknown points from curves rather than straight lines between the known points. These will not be considered here, Davis (2002) gives an extensive account of these methods, and this blogpost explains spline interpolation specifically.

Longer measurement time series in the Earth sciences covering a hundred year or more are very important to detect changes in the environment or climate. Much of the scientific debate on climate change is about the quality of the time series on which detection of climate change is based. Weather data have been gathered at some locations since the 18th century. During that time, many different observers have gathered the data, and technological developments have changed instrumentation. Changes in instrumentation especially may cause systematic changes in the observations. For instance, rainfall observations are very sensitive to the design of the rain gauge. On shorter time scales problems may also occur, such as instrument drift (small changes in the output of an instrument). In general, most time series need correction for these errors. Things get even more complicated if we want to combine meteorological measurements with reconstructions of past climate to extend our record of how climate has changed over time. Luckily, many talented data scientists work in the field of climate science, and the types of large datasets that are produced these days by institutes and consortia like NOAA, Copernicus, KNMI and Pages2k and which are used to inform reports like those by the IPCC are checked and double-checked by many scientists for issues like these.

V.2. Signal and noise: models of time series.

Time series analysis aims to find any mathematical relation between the observations in a time series. This relation may tell us something about the processes by which a time series is generated. To better understand time series analysis it is helpful to say something more about statistical models of time series. The simplest model is that of a time series consisting of random numbers. In such a time series, each observation is completely independent from the foregoing observations (figure V.3, top). The example in figure V.2 has been generated using the random number generator of the computer. When a correlation coefficient is calculated between all pairs of neighbouring values, so between all x_t and x_{t+1}, it would be approximately zero (in the example of figure V.2 it is 0.025). This type of time series is also known as 'white noise' - a peculiar name, but if you would convert the time series into sound it would sound like the hissing noise of a radio which is not tuned in on any station.

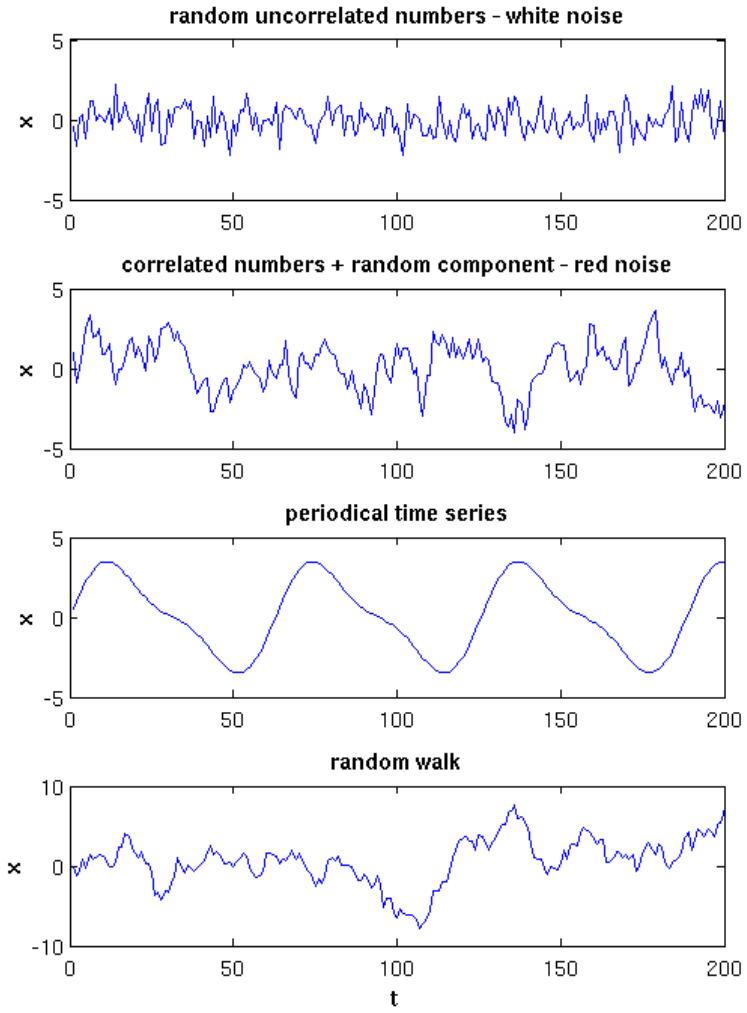

Figure V.3. Time series made using different mathematical models. Top: pure uncorrelated random numbers. Second: correlated, with random component based on the equation $x_t=0.8x_{t-1}+e$; third: periodical time series, generated by the equation $x_t= 3\sin(t/10)+\sin(t/5)$; bottom: 'random walk' generated by adding a random number to the previous value, $x_t=x_{t-1}+e$.

The second time series has more structure. Successive values look more the same, when one x_t has a low value, the next one, x_{t+1}, is likely to have a low value also. This time series is generated by the equation $x_t = 0.8x_{t-1} + e$, where e is a random number (error term). So, now the successive values in the time series are dependent on each other, and if you would calculate the correlation between all x_t and x_{t+1}, it would be significantly different from zero - in the example it is 0.788. This type of time series is called autoregressive, the relation between successive values can be determined by calculating a regression of the time series on itself, on x_t and x_{t+1}. It is also known as 'red noise'. It is also a kind of time series that often occurs in the earth sciences - although the example is generated purely artificially, it has a superficial resemblance to certain climate time series.

The third one in figure V.3 is a purely periodical one. The example is generated by the equation

$x_t=3sin(t/10) + sin(t/5)$, a summation of two sine waves with a different amplitude (3 and 1) and different periods (t/10 and t/5). A time series is a periodic time series when it satisfies the following equation:

$$x_t = x_{t+k} \qquad \text{V.2}$$

where k is a constant. It says that after every k timesteps, the same value of x returns. Many climate time series, eg. the glacial-interglacial cycles of the Quaternary, have a periodic component. In fact, these time series often consist of periodic components with added autoregressive components or 'red noise'. In the last section of this chapter we will learn how to detect periodic components, the length of their periods and their amplitude.

These three time series models are used very often in time series analysis. They have one property in common: they are bounded - their values will never exceed a certain maximum or minimum. The last one in figure V.3 does not adhere to this property. It is made by adding a random number to the previous value: $x_t = x_{t-1} + e$, where e is the random number. As if you would make a walk by throwing a dice to determine how many steps in one or the other direction you would take - therefore it is known as the random walk.

A way to recognize which model of time series applies is the *autocorrelation function*. We have already seen that calculating the correlation between all x_t and x_{t+1}, shows whether successive values in the time series are dependent on each other. We can do this also for x's that are more than one step from each other, for instance k steps, so the correlation between x_t and x_{t+k}. The constant k is also known as the *lag* (from 'lagging behind') for which the correlation is calculated. The formula for autocorrelation function of a finite time series of length n an lag k is defined as

$$r_k = \frac{n-k \sum_{t=1}^{n-k} x_t x_{t-k} - \sum_{t=1}^{n-k} x_t \sum_{t=1}^{n-k} x_{t-k}}{\sqrt{\left[n-k \sum_{t=1}^{n-k} x_t^2 - \left(\sum_{t=1}^{n-k} x_t\right)^2\right]\left[n-k \sum_{t=1}^{n-k} x_{t-1}^2 - \left(\sum_{t=1}^{n-k} x_{t-1}\right)^2\right]}} \qquad \text{V.3}$$

When $k=0$, r_k is equal to the standard deviation of x_t. The shift of the time series can be in a positive or negative direction, k can be positive or negative. The graph of r_k is symmetrical around $k=0$, since at both a positive and negative shift the same pieces of the time series are correlated with each other. The autocorrelation values are usually graphed with the values of k on the horizontal axis, and r on the vertical axis.

The shape of the autocorrelation function is a characteristic one for different time series models. For the first three time series of figure V.3, the autocorrelation function is shown in figure V.4. The 'white noise' has a correlation of 1 at $k=0$, and near-zero correlations at $k\neq0$. This shows, that successive values are uncorrelated at all lags besides 0 - where the correlation is of course perfect since we correlate all x_t with themselves. The autoregressive time series has high correlations at lags close to 0. The correlations drop gradually to near-zero values with larger absolute values of k. The

larger the distance between two successive values, the smaller the influence of previous values on the next ones. The autocorrelation function of the periodic time series is also a periodic function, since by definition the values of the time series are equal to each other at multiples of the value of *k*, that equals the period of the time series.

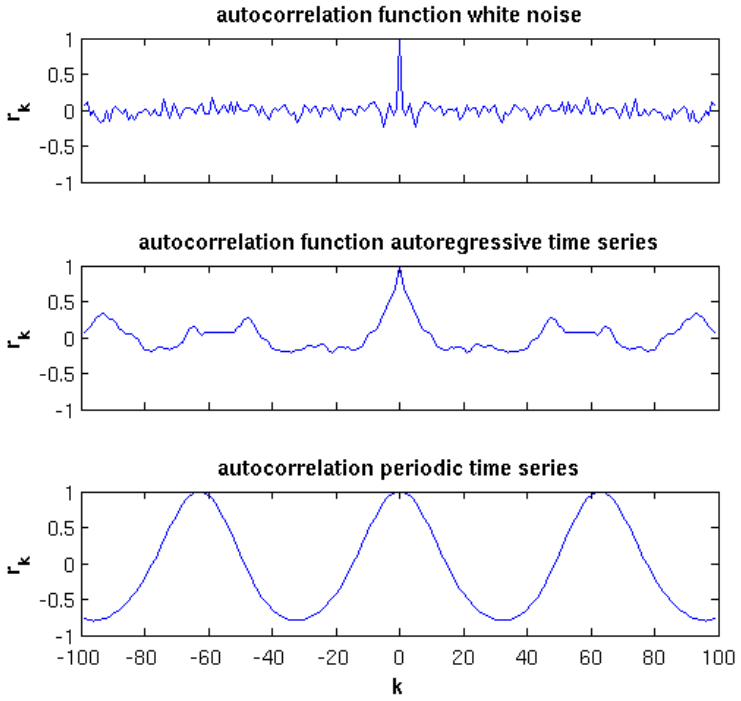

Figure V.4. Autocorrelation functions of the first three time series of figure V.3. Top: white noise (randum, uncorrelated) time series; middle: autoregressive time series; bottom: periodic time series. On the horizontal axis the lag (k), on the vertical axis the correlation between the time series and itself.

Most 'real life' time series are mixtures of these models, they may contain periodic, autoregressive and random components. The random components or *noise* may mask the *signal* of the process that generated the time series, as shown in figure V.5. This figure shows a sine wave of decreasing amplitude. At the left side of the graph, the amplitude of the signal is large enough to overcome the noise. To the right side, the amplitude has become smaller than the amplitude or power of the noise, and the signal is completely masked by the noise.

Time series analysis contains many methods to isolate these periodic or autoregressive parts, and to reduce the random noise. A simple method for an equal time step time series is '*smoothing*' the time series by taking a *moving average*. For calculating a moving average, an uneven number of consecutive values is taken from the time series, a 'window'. Then, the average of these values are taken and assigned to the middle value of the time series. Let *y* be the smoothed version of time series x_t, and the window length be *2n+1,* then *y* is defined by:

$$y_t = \frac{\sum_{t-n}^{t} x_t}{2n+1} \qquad \text{V.4}$$

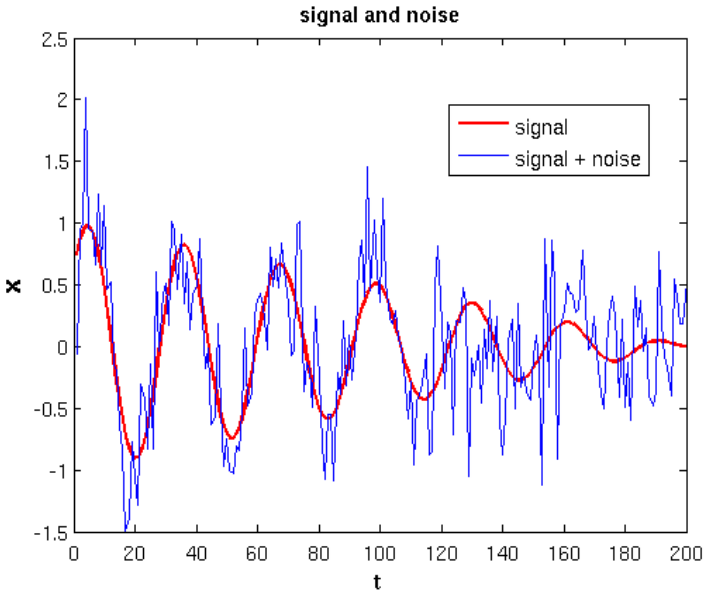

Figure V.5. 'Noise' masking the 'signal' in a time series. The signal is a sine wave of decreasing amplitude. Superposed on this is random fluctuation, the noise. To the left side of the graph, the signal is strong enough to be distinguished from the noise. To the right, the signal is not recognizable from the noise.

This type of treatment is also known as a 'filter', in this case a smoothing filter. Several other types of filters can be constructed to highlight features of time series.

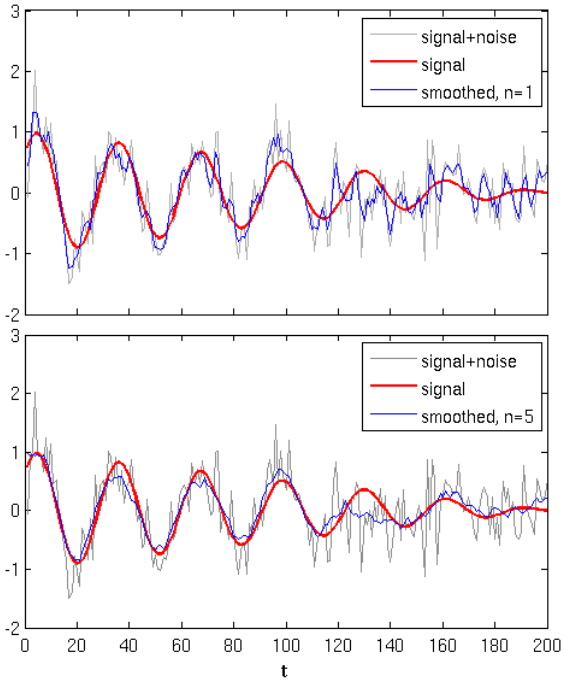

Figure V.6. Moving average windows applied to the noisy signal of figure V.5. Top: window with n=1, (window length 3 time steps), bottom: window with n=5 (window length 11 time steps)

In figure V.6, smoothing windows have been applied to reduce the noise in the time series of figure V.5. The top graph of figure V.6 shows a window of 3 time steps long (*n*=1). The resulting time series still shows a considerable part of the random noise, although it follows the original signal more closely. To the right part of the graph, the signal remains invisible. In the bottom graph a window of 11 time steps is used (*n*=5). Here, the resulting smoothed time series more strongly resembles the signal, even in the rightmost part. Clearly, the longer the window, the better results. Bu there is an upper limit: if we would make the window longer than half the period of the time series (here 50 time steps), also the signal would have been smoothed away. So the choice of the window length also depends on knowledge of the signal.

Another property of time series is their evolution in time. If a time series is divided into smaller segments and the means of these segments is the same everywhere and the same as the mean of the entire series, it is said to be *first order stationary* (figure V.7). If the same holds for the standard deviation of these segments, is *second order stationary*. If first-order stationarity does not apply the time series is evolutionary. It may display a regular trend for instance.

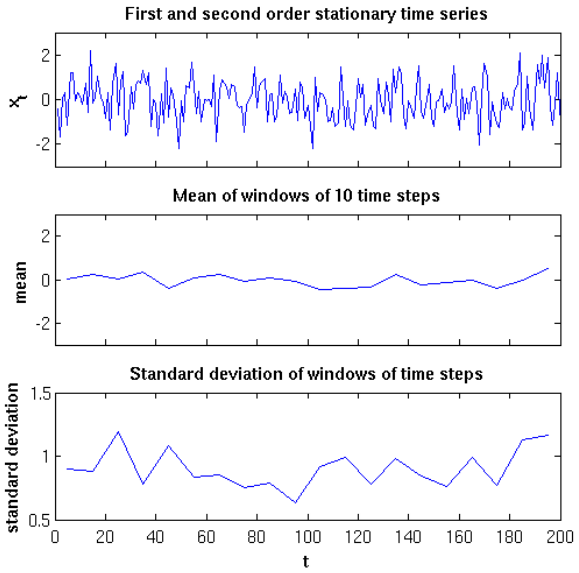

Figure V.7. The white noise time series of figure V.4 is a stationary time series (top). Middle: mean of 10 time step long windows. Bottom: standard deviation of the same windows.

For many types of analysis it is useful to to remove any trends in the data, e.g. for spectral analysis in the next section. A way to remove this trend is to calculate a regression line with t as independent variable and x_t as dependent variable, and to subtract the regression line from the data. Once the equation of the regression line has been determined, the value of the trend can be calculated from it for every t, and subtracted from x_t. At the same time, the ANOVA table of this regression indicates whether the trend is significant or not, using the procedures described in chapter I. If necessary polynomial regression also can be used to remove trends that are not straight lines. Figure V.8 shows an example for a linear trend.

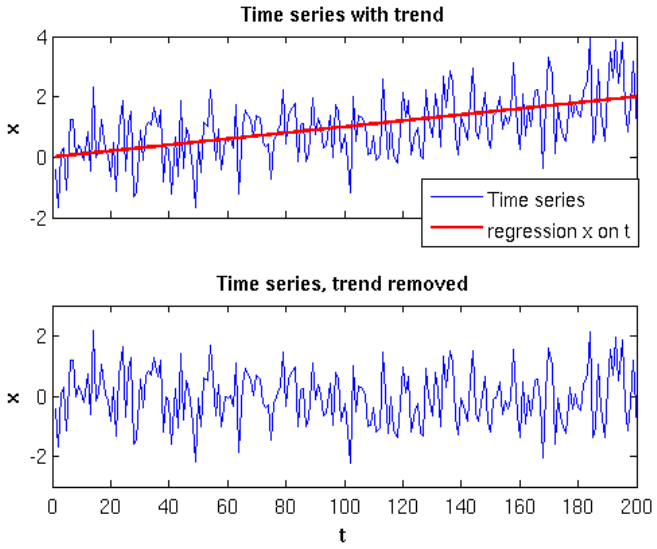

Figure V.8. Time series with a linear trend. The trend is removed by calculating the regression of x on t, and subtracting it from x.

V.3. Periodicity.

In the previous section we have seen that a time series may contain a periodic signal. Such periodic signals are very common in the Earth Sciences. Think of the glacial-interglacial cycles of the Quaternary, and similar cycles in older eras, that are driven by the Milankovich cycles. Also, daily and seasonal cycles commonly occur. Techniques to detect these cycles in otherwise noisy records have been applied very frequently, and any self-respecting earth scientist should have a basic knowledge of these techniques. In fact, the very link between ice ages and Milankovich cycles has been confirmed by the spectral analysis methods discussed below.

Spectral analysis is a technique that isolates periodic components from a time series, and quantifies these components using their basic characteristics: frequency / wavelength, amplitude / power, and eventually phase. In time series, several periodic components may be present, with different wavelengths. For instance in a Quaternary climate time series you may expect to find the ~100.000 year (100 kilo-year, kyr) glacial-interglacial cycle, a ~40 kyr cycle and ~20 kyr cycles, each related to one of the Milankovich cycles.

Figure V.9 gives a short overview of what wavelength, amplitude and phase angle are.

Amplitude (*A*) is the magnitude of maximal deviation from the mean of a periodic component (figure V.9). A change of amplitude occurs by multiplying the function with a constant. It is a measure of how strong a periodic signal is present in a time series. The amplitude is a measure of the power of a periodic component of the time series. As we have seen in figure V.6, the amplitude determines whether a signal can be distinguished from random noise.

Wavelength (λ) is the same as the period length of the periodic time series in the previous section. It can also be expressed as frequency, the number of cycles per time unit. For instance, sound frequency is usually measured in cycles per second, ranging for us humans from 20 to 20.000 cycles per second or 20.000 Hertz (abbreviated Hz). It will be clear that in geology longer time units may be used - years or the kiloyears above. The relation between frequency f and wavelength λ (lambda) is reciprocal:

$$f = 1/\lambda \qquad \text{V.5}$$

In a trigonometric function e.g. *sin(t)*, the wavelength changes when the horizontal axis values are multiplied by a constant, e.g. *sin(2t)* doubles the frequency and *sin(0.5t)* halves the frequency (figure V.9, third graph).

The *phase angle (ϑ)* determines a horizontal shift of a trigonometric function to the left or right (figure V.9 bottom). The phase changes when a constant is added to the horizontal axis values.

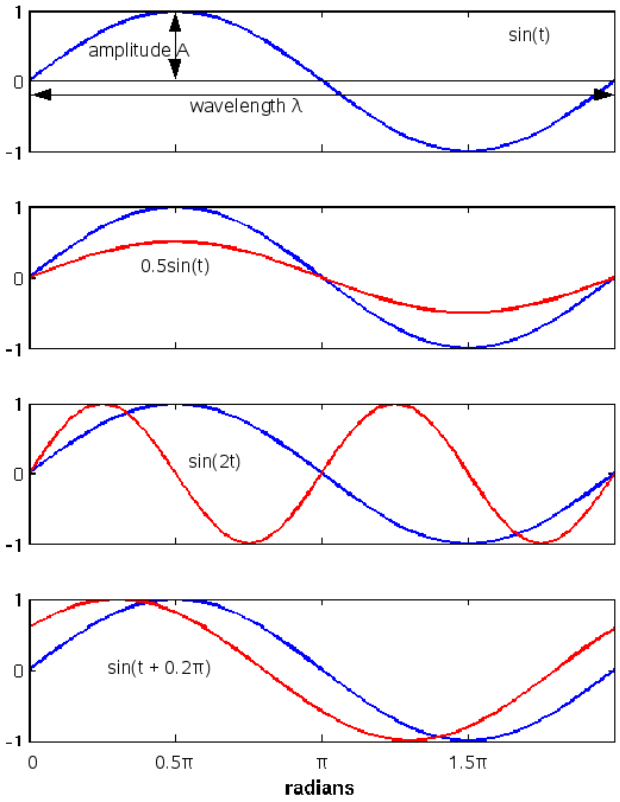

Figure V.9. Top: sine function, showing amplitude and wavelength. Second graph: multiplying with a constant increases or decreases the amplitude. Third: multiplying the time scale with a constant increases or decreases wavelength / frequency. Bottom: phase shift by adding a constant to the time scale.

In summary, a periodic time series consisting of a single periodic component may be characterized by an equation of this type:

$x_t = A \sin(kt + \vartheta)$ or $x_t = A \cos(kt + \vartheta)$ V.6

where A is the amplitude factor, k the wavelength factor and ϑ the phase angle. Note that the difference between the sine and the cosine is a phase shift of ½π.

Spectral analysis is used to detect periodic signals of different frequency in time series, and to determine their amplitude. The results of spectral analysis are usually displayed in the shape of a *spectrogram.* On the horizontal axis of a spectrogram appears wavelength or frequency, on the vertical axis power or amplitude. So from the spectrogram you can read what the wavelength is of the periodic components, and what their power is. Figure V.10 (right side) shows an example of a spectrogram.

In figure V.10 left, the original time series is displayed. It is a grainsize record from a thick pre-Quaternary loess section in central China. In these loess deposits, consisting of windblown dust from the central Asian deserts, climate changes are recorded by the grainsize of the material. Colder, more windy climate caused deposition of coarser grains while in warmer and wetter climates finer grains were deposited. The time series has a vertical time axis (measured in kiloyears) and a horizontal grainsize value axis.

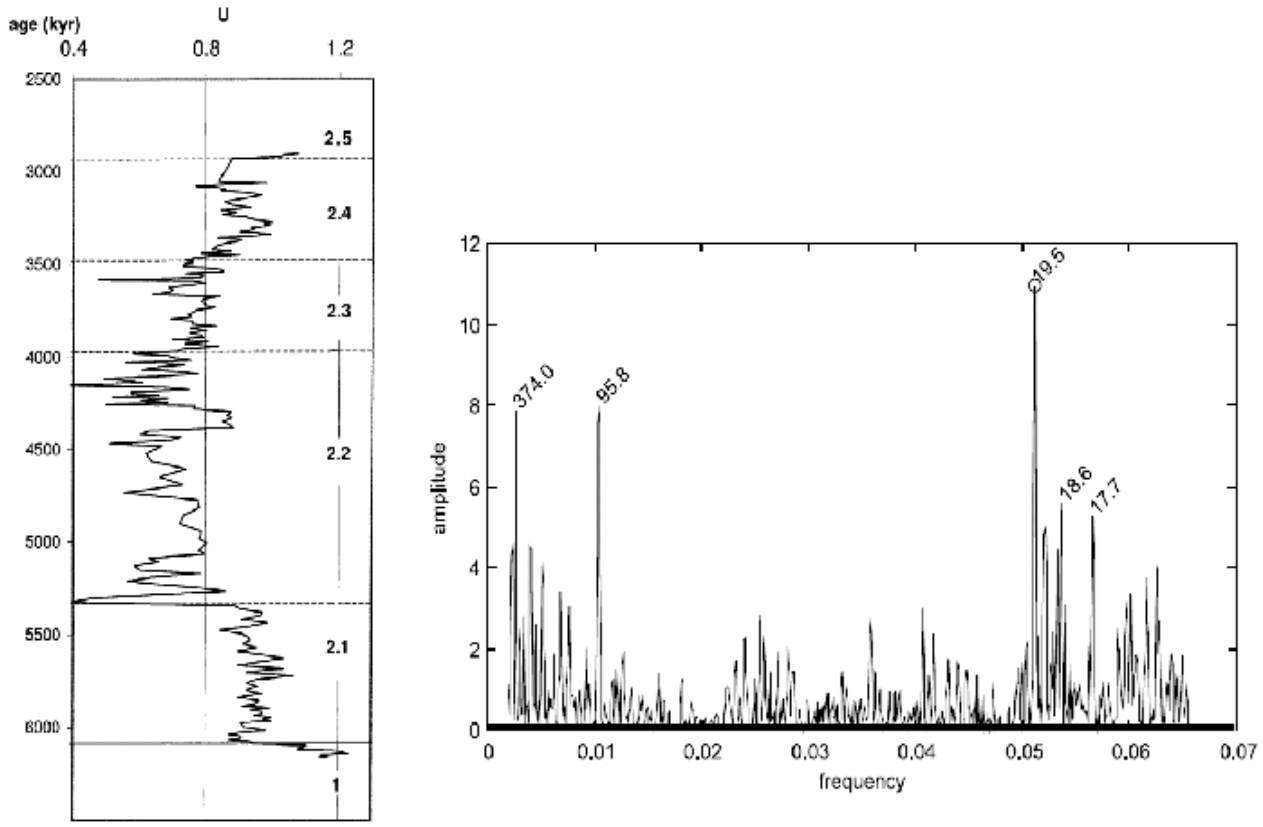

Figure V.10. Left: time series derived from a geological section in Chinese loess deposits. On the vertical axis time in units of thousand year (kiloyear), horizontal axis: a grainsize parameter. Right: spectrogram of the same time series. On the horizontal axis the frequency, in cycles per thousand year, on the vertical axis, the amplitude. The peaks in the spectrogram are clear periodical components of the time series. At the highest peaks also the wavelength (in kiloyear) is given.

On the right side the spectrogram derived from the time series is shown. The horizontal axis of the spectrogram is the *frequency*, the vertical axis the *amplitude* - quite different from the time series! The graph of the spectrogram is a rather irregular collection of smaller and larger peaks. Each of these represent a periodic component, the size of the peak denotes the amplitude. You might conclude that there are many, many periodic components in the time series, since there are many peaks. However, only the larger ones are probably significant, the smaller ones could have arisen from noise/errors in the time series. Noise results often in many high frequencies with a small amplitude in spectral analysis (see below). The frequency of the horizontal axis is given in cycles per thousand years, which is somewhat difficult to interpret. Therefore for the larger peaks the frequency has been converted to wavelength in kiloyears using formula V.5.

Having seen the results of spectral analysis we should consider how these results are obtained. Like with factor analysis, there are many methods, each with their own mathematics, assumptions and terminology. However, most methods are based on the Fourier transform. Every periodic function can be expressed as the sum of an infinite array of sine and cosine functions:

$$x_t = \sum_{k=1}^{\infty} [\ _k\cos\ k \qquad _k\sin\ k\ ] \qquad \text{V.7.}$$

Each of these functions have their on frequency, determined by *k= 1, 2, 3...∞. k* has only positive integer numbers, it is known as the *harmonic number*. For every k, the sine and cosine functions have a specific amplitude, α_k and β_k. A graphic example is shown in figure V.11.

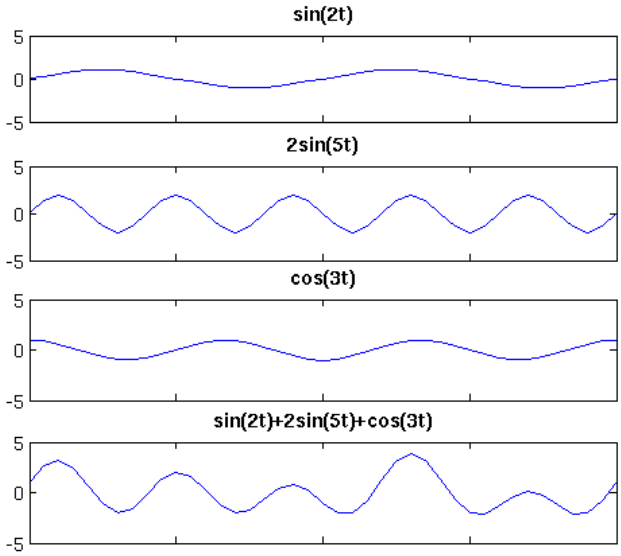

Figure V.11. Summation of sine and cosine functions with different amplitudes and wavelengths.

We can also apply this to a time series, on the assumption that it is a periodic function. However, the angle ϑ in equation V.7 is given in radians, not in units of time as in time series. Still, the time scale of of a time series also can be converted into radians, by converting *t* into a fraction of the total length *T* and multiplying by 2π:

$$_t = \frac{2\ t}{T} \qquad \text{V.8a}$$

or, in the case of an equally spaced time series with *n* observations

$$_t = \frac{2\ t}{n} \qquad \text{V.8b}$$

So, V.7 then changes into:

$$x_t = \sum_{k=1}^{\infty} \left[\alpha_k \cos\frac{2\pi kt}{n} + \beta_k \sin\frac{2\pi kt}{n} \right] \qquad \text{V.9}$$

The next task is then to find the amplitudes for every harmonic number *k*. This can be done by transforming V.9 into a regression equation:

$$x_t = \alpha_0 + \sum_{k=1}^{n-1\ /2} \left[\alpha_k \cos\frac{2\pi kt}{n} + \beta_k \sin\frac{2\pi kt}{n} \right] \qquad \text{V.10}$$

The α's and β's are then regression coefficients to be estimated. How this is done, will not be discussed here. The main point is that the coefficients can be estimated from the time series data:

$$\alpha_k = \frac{2}{n}\sum_{t=1}^{n} x_t \sin\frac{2\pi kt}{n} \quad \text{and} \quad \beta_k = \frac{2}{n}\sum_{t=1}^{n} x_t \cos\frac{2\pi kt}{n} \qquad \text{V.11}$$

From these two coefficients we can determine the amplitude of the periodic component with frequency determined by *k*:

$$A_k = \sqrt{\alpha_k^2 + \beta_k^2} \qquad \text{V.12}$$

The α_0 constant is determined by

$$\alpha_0 = \frac{1}{n}\sum_{t=1}^{n-1} x_t \qquad \text{V.13}$$

which is just the mean of the time series.

After all the computational work - not to be done by hand calculator but by computer please - the spectrogram is obtained by plotting the amplitude A_k against harmonic number *k*. In general, *k* is converted to frequency.

Figure V.12. Spectrograms of the upper three time series of figure V.4. The insets show details of the spectrograms.

Figure V.12 shows example spectrograms of the upper three time series of figure V.4. All

spectrograms have been plotted on the same scale. The insets to the right of the spectrograms show details on a larger scale.

The topmost spectrogram is from the periodical function. The equation of this function is $x_t = 3\sin(t/10)+\sin(t/5)$. From this equation we can see that it should have two periodic components, the first with half the wavelength or double frequency of the second, and a three times higher amplitude of 3 units. Indeed we can see these components in the spectrogram as two separate peaks. The first peak occurs at a frequency of 0.0156, or a wavelength of 1/0.0156 = 64.10, and an amplitude of 2.76. The second one has a frequency of 0.0312, wavelength of 32.05, and amplitude 0.35. In figure V.4 it is easily checked that the wavelength of the strongest sine wave should be around 64.10. The wavelength of the second sine is half that of the first one, as expected. Only the amplitudes appear smaller, in particular for the second sine - this should be 1.0 instead of 0.35. This needs some explanation. The reason is, that the Fourier transform above gives only estimates at discrete harmonic numbers *k*. These may not match the frequencies in the data exactly, in particular for low frequencies on the left side of the spectrogram. If there is not an exact match this may cause a lower amplitude estimate.

The next three spectrograms show the effects of noise. The second spectrogram is from the white noise time series in figure V.4. No peaks are visible, for all frequencies the amplitude is nearly zero. Only when the spectrogram is strongly magnified as in the inset, we can see an irregular pattern of peaks. The third spectrogram is from the autoregressive time series in figure V.4. It shows modest peaks at low frequencies, none very dominating. This pattern is common for this type of time series. In the bottom spectrogram a combination of the red noise and periodic time series has been made. This type of mixed periodic + autoregressive noise time series is very common in earth sciences. In the spectrogram we can clearly see the 0.0156 frequency peak of the periodic time series. However, its second peak cannot be distinguished now from the peaks caused by the noise.

These examples show how we can interpret spectrograms. However, in real life Earth science time series it can be much more difficult to decide whether a peak in a spectrogram is caused by random noise, or by a real periodic signal. Therefore, most spectral analysis methods provide significance tests for helping with this decision.

Spectral analysis methods are generally also more intricate than the simple Fourier spectrum shown here. This has its origin in the fact that the Fourier transform is meant for infinite periodic time series, while our time series are neither infinite nor truly periodic, since they always contain noise. For this reason the analysis is not applied to the time series directly, but to its autocorrelation function. The highest frequency that can be detected by the Fourier transform has a harmonic number *k* of $(n-1)/2$. This is called the Nyquist frequency. Its wavelength is equal to 2 times the time step of the time series. Unfortunately, if higher frequencies are present, these can cause artificial lower frequencies to be present in the spectrogram. To reduce these effects, most spectral analysis method incorporate other mathematical manipulations, such applying filters that remove the highest frequencies to the data. See Davis (2002) for more explanation.

V.4. Extreme values in time series.

An important question for many Earth science time series is the occurrence of extreme events. These extreme events are nearly always the disasters that cause loss of life, property and otherwise serious damage to ecosystems and society. The extreme events are the earthquakes, storms and floods that make insurance companies nervous, and mercilessly lay bare the weakness of even the most powerful governments, as hurricane Katrina did show in 2005.

When talking about risks, government agencies and the media often quote return times or *recurrence intervals* to express how often an event occurs. For instance Dutch Rijkswaterstaat says that the dikes in the Netherlands should resist events with a recurrence interval of once in 4000 years or even once in 10000 years. This section serves to explore what these figures mean and how they are obtained.

A recurrence interval is based on how many times a certain event occurs in an observational record, and can be derived using these simple formulae:

$$T = N/n \qquad \text{V.14}$$

where N is the number of years of the record, and n is the number of events. For events that have a certain magnitude, such as river floods, the following formula applies:

$$T=(N+1)/m \qquad \text{V.15}$$

where m is the rank of the event. To determine m, the events in the record have to be ranked according to size. The event of highest magnitude gets the lowest rank. E.g. in a river discharge record of four years, 311, 520, 250, 756 m^3/s, the discharges will be ranked in the order 756, 520, 311, 250 m^3/s, with rank 1, 2, 3 and 4. The discharge of 520 m^3/s then has a recurrence interval T of (4+1)/3 = 5/3 = 1.67 year.

The recurrence interval is closely related to the probability that an event is exceeded in a year:

$$P(x_t \geq X) = \frac{1}{T} \qquad \text{V.16}$$

If the chance of occurrence of a flood larger than value X is, say, 0.01, than its recurrence interval is 1/0.01 = 100 years.

Two remarks here.

First, the interpretation of *T* is that of a probability. A recurrence interval of 10 years for a rainfall of a certain magnitude does not mean, that once this event has happened, it will take another ten years before it happens. It may happen the next year, or the next month again. The formulas above do not say anything about the distribution of the events - these may have occurred in a cluster, or randomly or evenly spread over the record.

Second, it will be clear that the longer the observational record is, the more reliable the estimate of *T* will be. However, observational records in Earth sciences are in general not that long, a few hundred years in most instances. These records can sometimes be extended by historical or by geological research. However, the farther back in time, the more uncertainties.

If we do not have records longer than a few hundred years, how can a government agency say that the dikes in the Netherlands should withstand a 1 in 10000 year event? In fact, there is a confusion in terminology here. What the government agency refers to, is not the recurrence interval based on a finite observation record in the formulas above. Instead, a probability of event magnitude is meant here. This probability is forecasted from the observations using statistical techniques. How this is done, is discussed below.

If we could obtain a probability distribution of the magnitude of events from a finite observation record, we could also estimate the probabilities for the more extreme events. A probability distribution that is useful for this purpose is the Gumbel distribution. It is a skewed distribution, and holds only for real numbers equal to or larger than 0. It is designed to find the probability of the maximum or minimum of a number of samples, and proves to do well in hydrology for estimating the probability of river discharge maxima.

The formula for the probability distribution *p* and cumulative distribution function *P* is

$$p(x;\mu,\alpha) = \frac{1}{\alpha} e^{\left[\frac{x-\mu}{\alpha} - e^{\frac{x-\mu}{\alpha}}\right]} = \frac{1}{\alpha}\exp\left[\frac{x-\mu}{\alpha} - \exp\left(\frac{x-\mu}{\alpha}\right)\right] \qquad \text{V.17}$$

$$P(x;\mu,\alpha) = e^{-e^{\frac{-x}{\alpha}}} = \exp\left[-\exp\left(\frac{-x}{\alpha}\right)\right]$$

The μ and α parameters determine the location along the x axis and the spread of the function, just like the mean and standard deviation of the normal distribution (figure V.13). And just like the mean and standard deviation, these parameters can be estimated from the data.

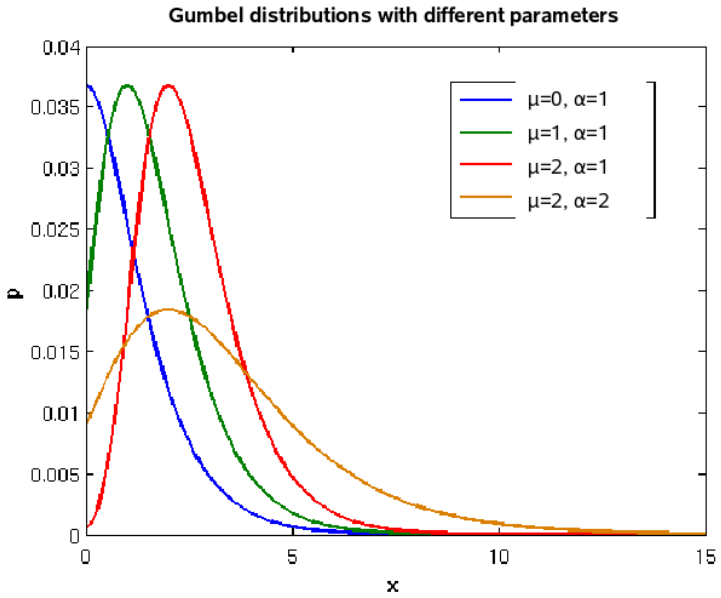

Figure V.13. Gumbel distribution. The effect of parameter μ on the location of the modus on the distribution and the parameter α on the width of the distribution.

Once we have estimated the parameters, we can determine the probability for events of any magnitude. However, keep in mind that this is just a statistical model, it is as good as the data and the assumptions on which it is based.

One assumption is that the Gumbel distribution is the right distribution. For high magnitude events, in particular the rightmost tail of the distribution is critical - very small differences there can change probabilities enormously. Other probability distributions exist with similar shape as the Gumbel, e.g. the Weibull distribution. These might fit the data set as a whole better, but give quite worse estimates for high magnitude events because of deviations in the distribution tail. A second assumption is, that the distribution represents one population of events. If the discharge regime of a river changes, for instance by climate change or by changes in land use and river management in the basin, the population essentially changes, causing a change of the probability distribution.

The parameters of can be estimated by calculating the normal mean $x$and standard deviation s of the data:

$$= 0.77987s$$
$$= x - 0.5772 \qquad \text{V.18}$$

Once the distribution parameters have been determined we can calculate the probability $P\ x_t \leq X$ for any event equal to or exceeding a certain limit X, and eventually convert it to a recurrence time by equation V.16.

Below an example. Figure V.14 shows a 12 year daily river discharge record from the Dinkel river on

the Dutch-German border. Winter discharges are highest, and peak discharges exceeding 30 m^3/s frequently happen although the discharge may drop as low as 1 m^3/s. For Dutch rivers, this river has a quite peaked discharge regime, due to its origin in a region underlain by impervious shales in Germany.

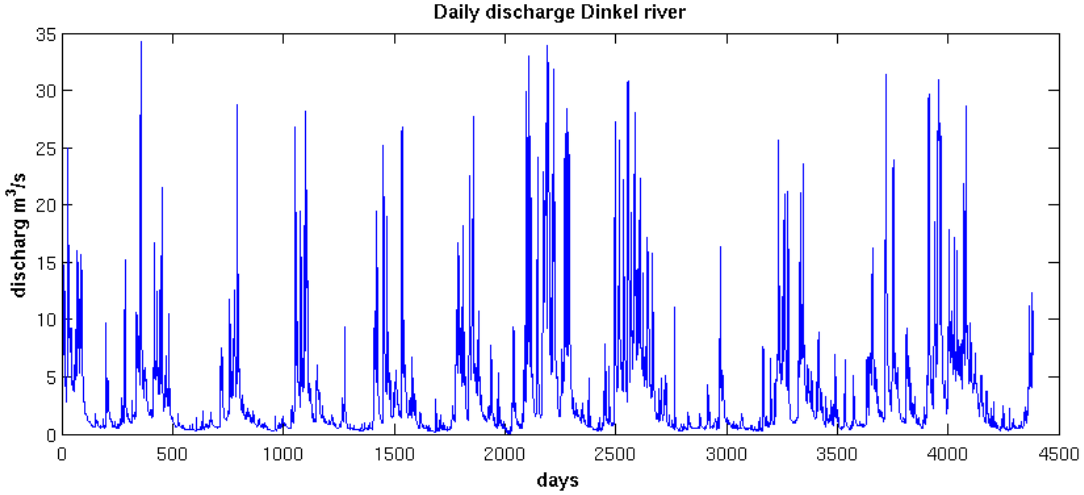

Figure V.14. Daily discharges from the Dinkel river in the eastern Netherlands.

The Dinkel valley is a popular tourist area. The owner of a restaurant along the river bank has seen his terrace and parking lot flooded a few times, and now his insurance company wants to know how often this can happen. These events occur at discharges exceeding 33 m^3/s. There are only very few of these events in the record: two in 12 years, suggesting a recurrence time of 6 years.

First, the population mean and standard deviation of the data is determined: x= 3.5661 and s = 4.8996 respectively. Using the formulas V.18, this gives values for $= 0.7797 \text{x} 4.8996 = 3.8202$ and $= 3.5661 - 0.5772 = 1.3610$. Figure V.15 shows the graph of the Gumbel distribution with these parameters, together with a frequency histogram of the data.

Next we can use V.18 to compute the probability for a flood larger than 33 m^3/s by setting x=33:

$$P(x_t \leq 33) = \exp\left[-\exp\left(-\frac{x - }{}\right)\right] = \exp\left[-\exp\left(-\frac{33 - 1.3610}{3.8202}\right)\right]$$
$$= \exp[-\exp(-8.2820)] = \exp[-0.00025304] = 0.99975$$

Since $P(x_t \leq 33)$ = 0.99975, $P(x_t > 33)$ = 1 - 0.99974699 = 0.00025301 (now, the numbers behind the comma are important!). This results in a recurrence interval of 3952.46 - in days, since the data are also given as daily discharges. Dividing by the number of days in the year, we get a recurrence interval of 10.82 years, clearly more than that of the initial estimate of 6 years.

For lower discharges, the estimates are likely to agree better. A larger than 30 m3/s discharge occurs 11 times in the record, indicating that is should occur with a recurrence interval of 1.09 year. Computing the recurrence interval using the Gumbel distribution results in a recurrence interval of

1.34 years.

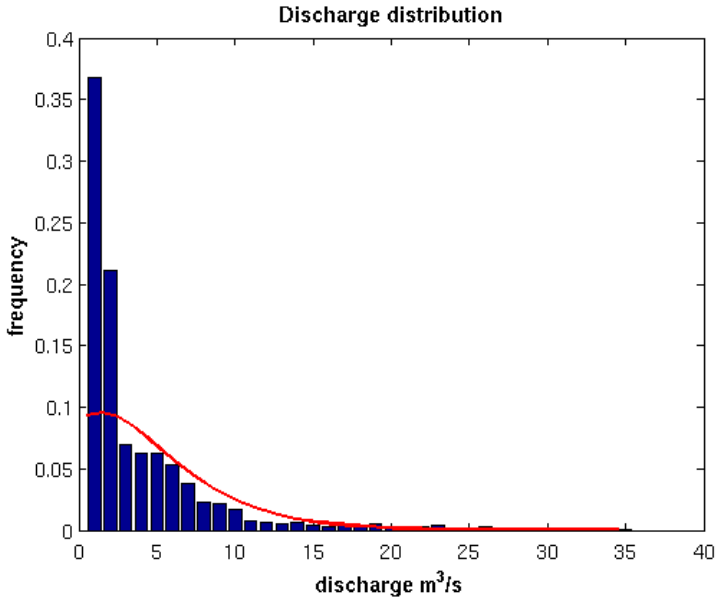

Figure V.15. Frequency histogram of observed discharges and the Gumbel distribution (red line) estimated from the data.