M.C. Thrun, AG Ultsch

# Knowledge discovery in big data time series - Hydrology

**Based on: Aubert, A. H., Thrun, M. C., Breuer, L., & Ultsch, A.: Knowledge discovery from data structure: hydrology versus biology controlled in-stream nitrate concentration, *Scientific reports, (in revision)*, 2016.**


Philipps Universität Marburg

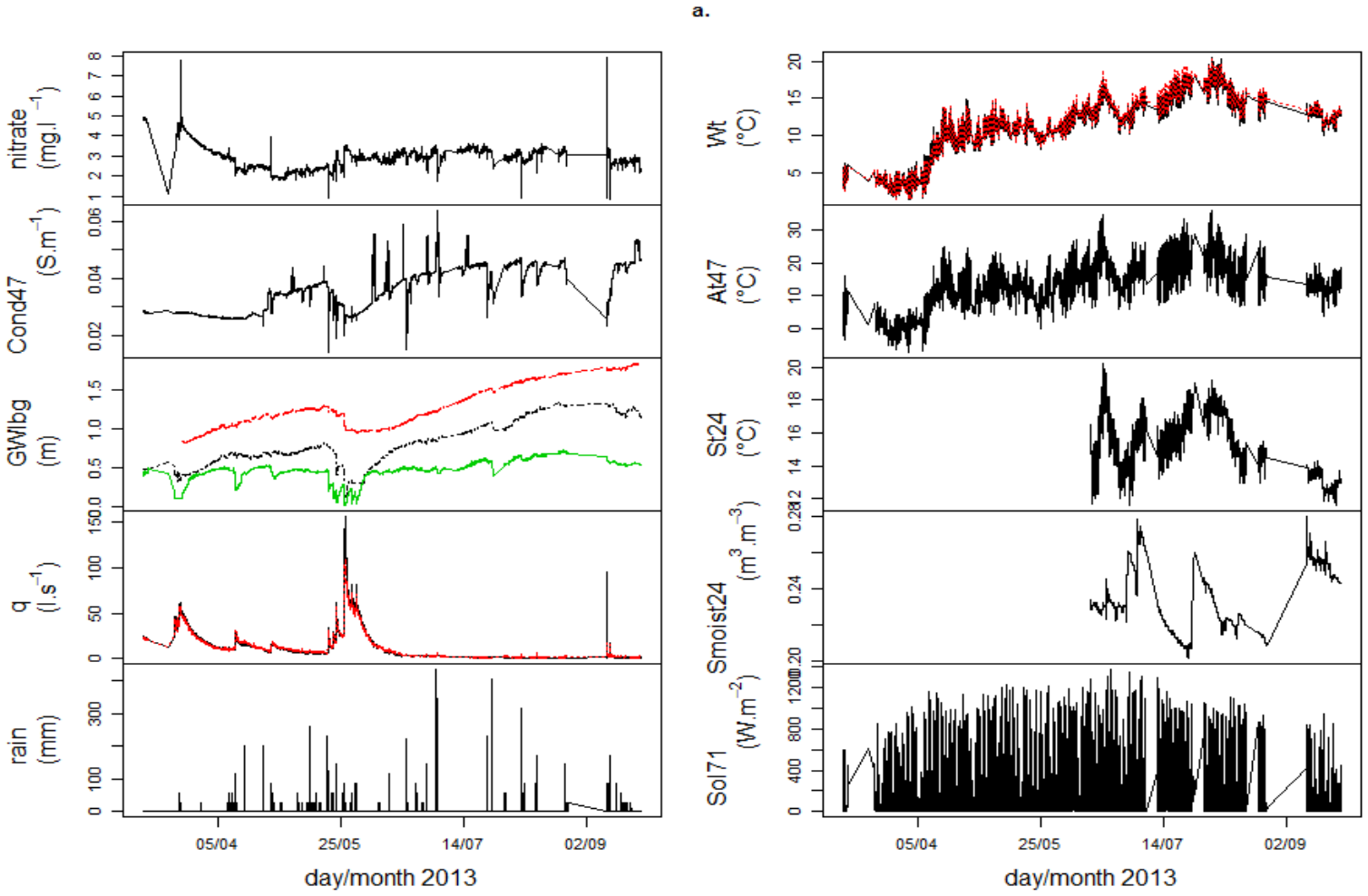
# Analysing a data set

- Goal: Mostly given by domain/topical experts
- Preprocessing: compound model using Fast Fourier Transformation (FFT)
- Analysis:
  - GaussianMixtureModell (GMM)
  - PDEplot
  - Statistics
- Interpretation by topical/domain experts

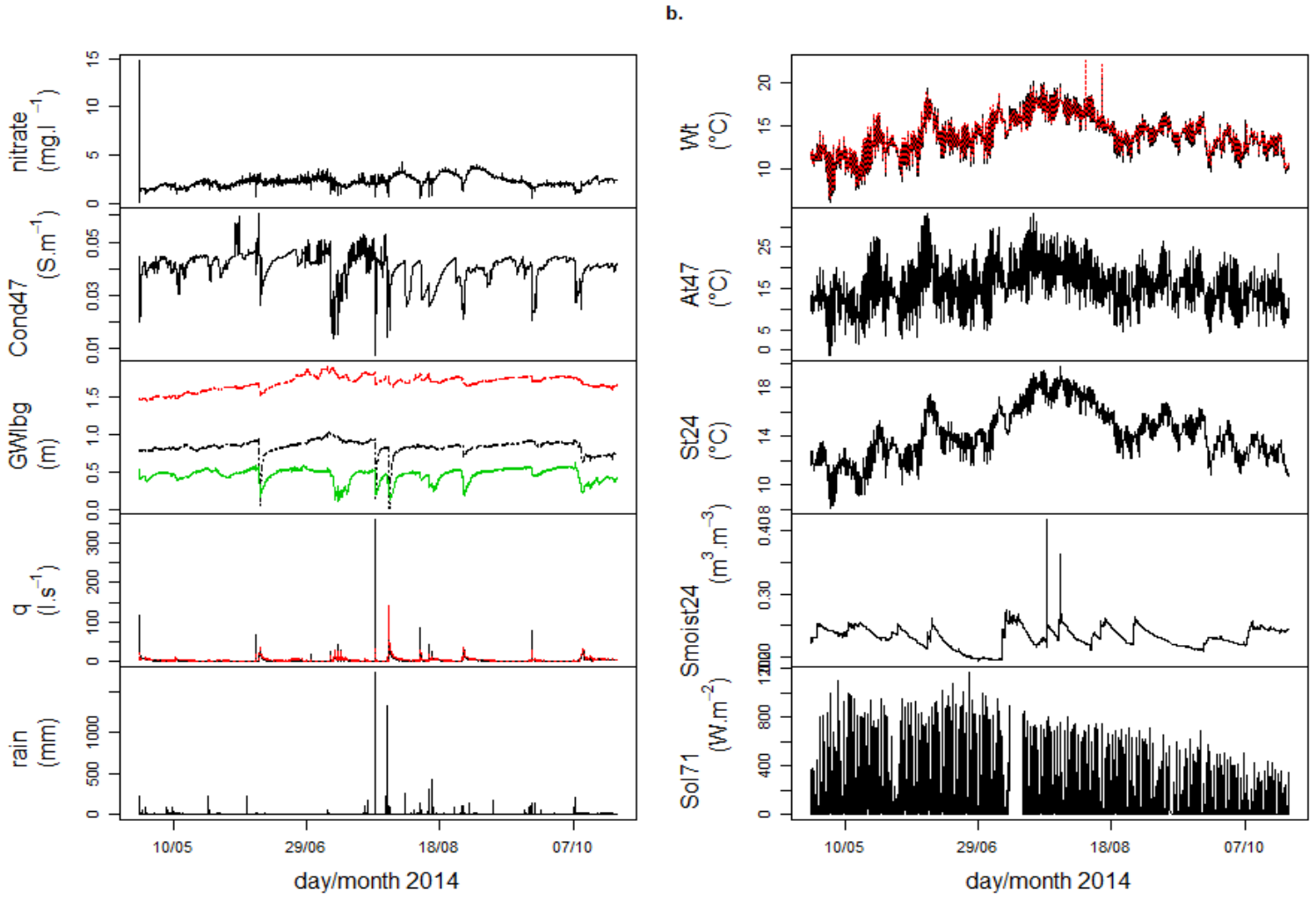
# Data set

- Environmental science
- High frequency measures in one area (~15min)
  - Catchment (~3.7 km²): ground and stream below ground
- Big Data: 2 years of measure points (>32 000)
- High-Dimensional (14 variables), e.g.
  - Solar radiation, Air temperature, Water temperature, Soil temperature, Soil moisture, Groundwater level, Discharge, Rainfall intensity, Electric conductivity, …

# Time series 2013



# Time series 2013



# Overview

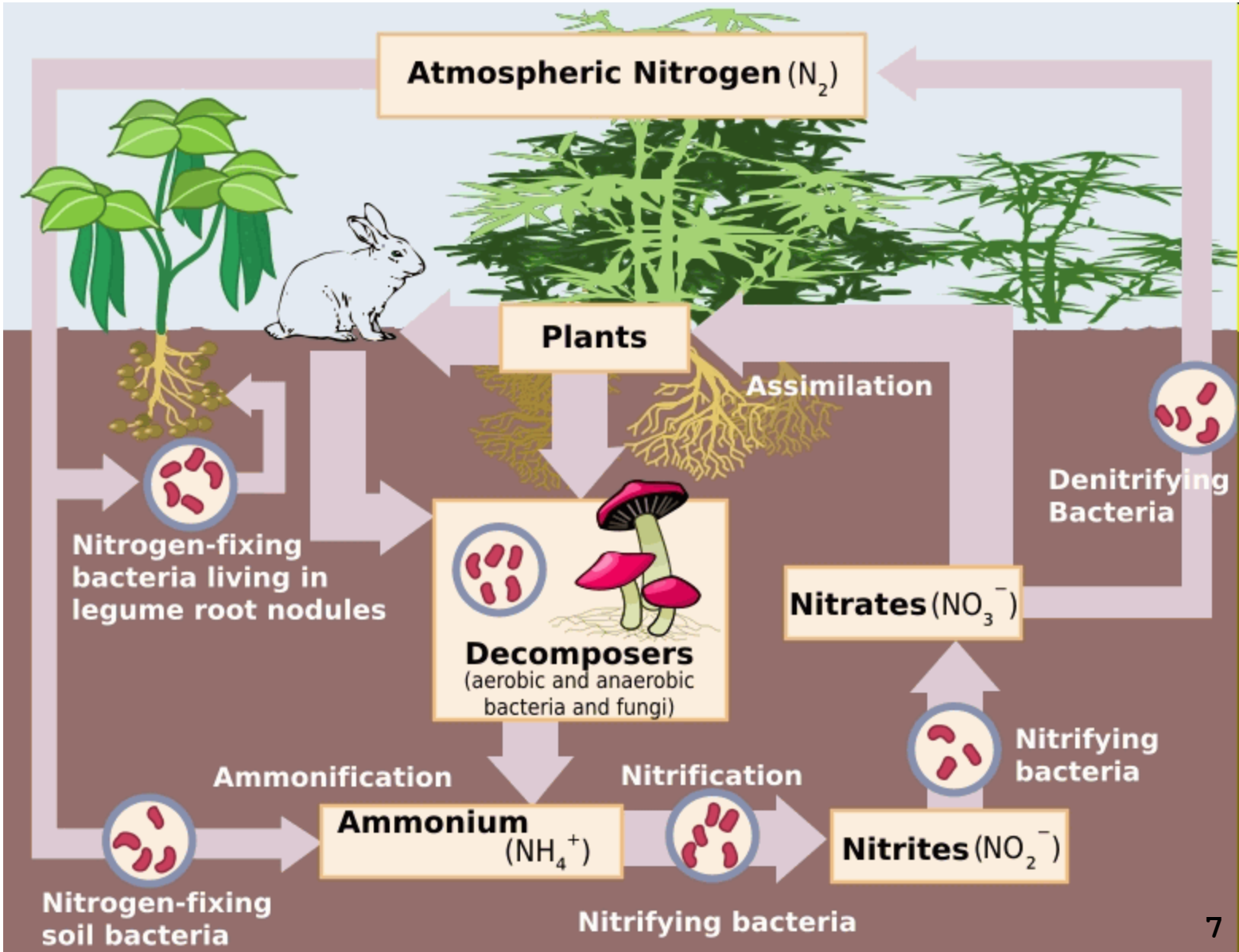
- Problem: Humans modify the nitrogen cycle, in particular by farming
  - For example impacts aquatic life
- What drives nitrate fluctuations?
- Are these drivers the same for low and high nitrate concentrations?

=> Approach: Remove temporality

# Expert‘s prior Knowledge: Nitrogen Cycle

**Source:**

**https://en.wikipedia.org/wiki/Nitrification**



# Multivariate time series preprocessing

- Detrend time serie to join both years
- Component model
  - Residuum=raw-Season-Trend
- Season: low pass filtering (50 days)
  - Fourier Transformation
- Residuum : Rapid high temporal fluctuation

# „Koch“-Rezept für Komponentenmodell

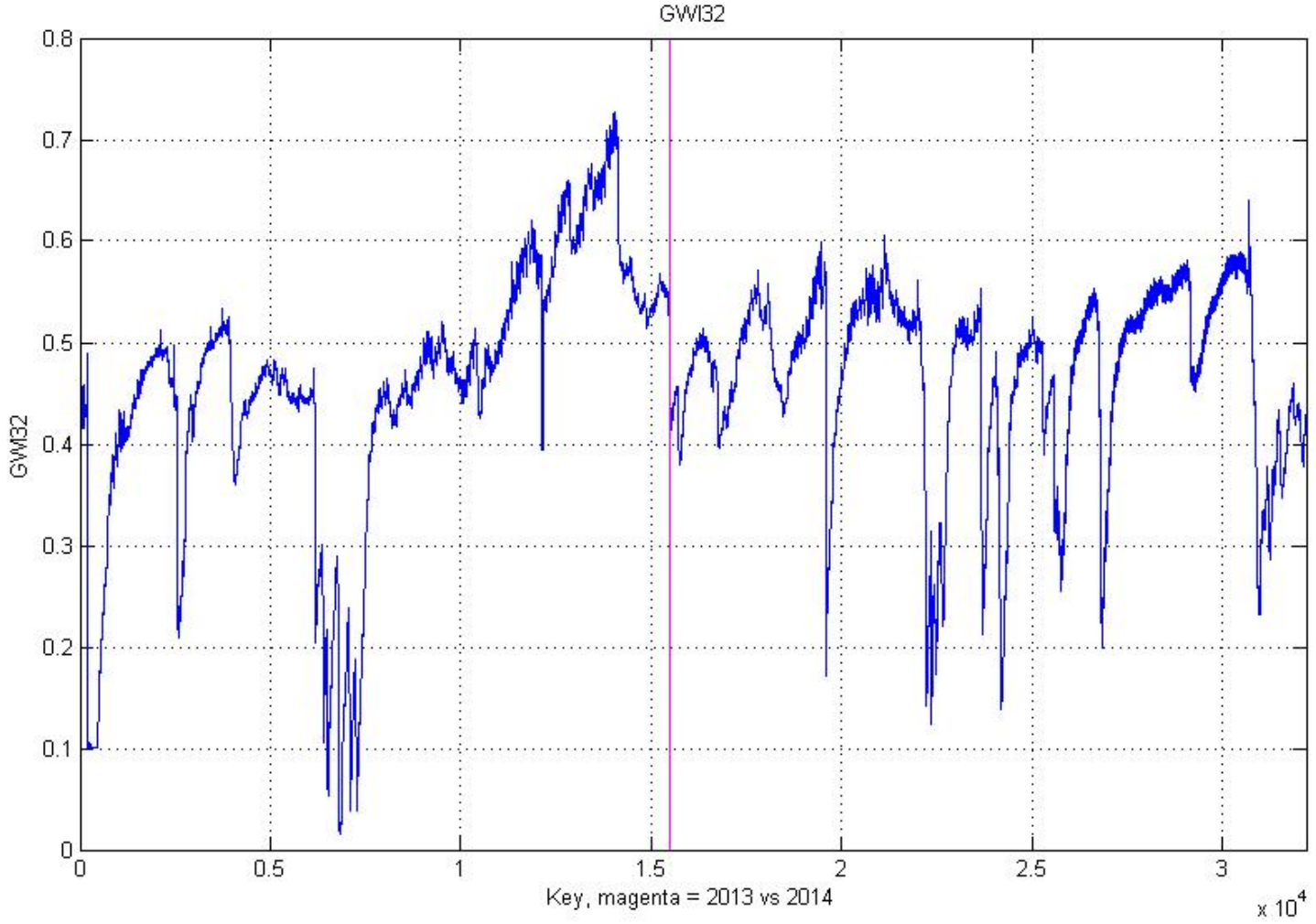
- Jahre getrennt pro Variable vorverarbeiten

1. Zentralen Mittelwert durch GMM abziehen
2. Zeitreihe spiegeln und verdoppelt
3. Fast Fourier Transformation (FFT)
4. Nyquist Frequenzen plotten
5. Grenzfrequenz für Tiefpassfilterung finden
6. Tiefpassfilterung: Restlichen Frequenzen entfernen
7. Zeitreihen für beide Jahre zusammenführen

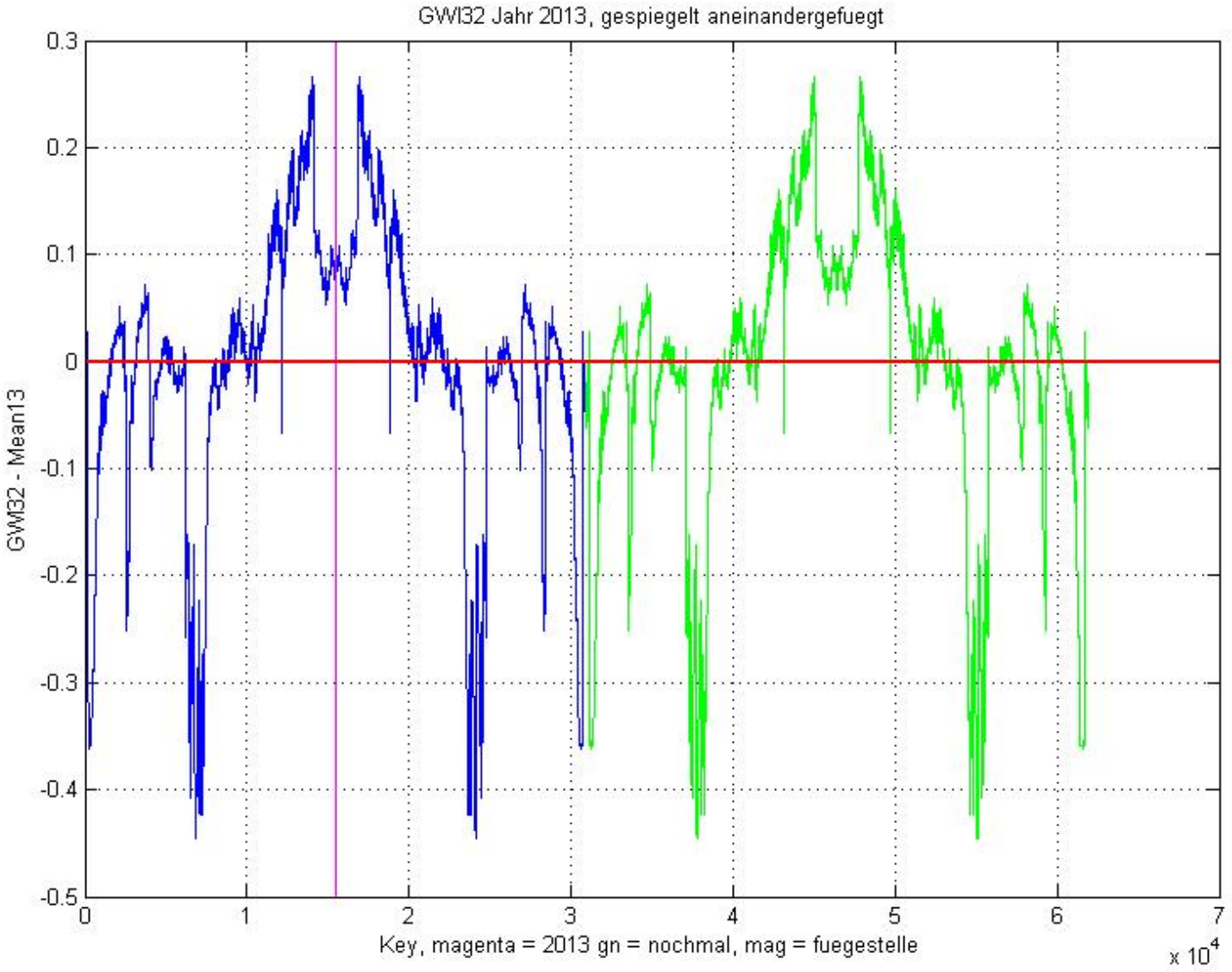
Keine Panik Erklärungen und Details kommen in VL!

# ZR für Jahre 2013 und 2014 weist Sprung auf

Groundwater level, 32=riparian zone

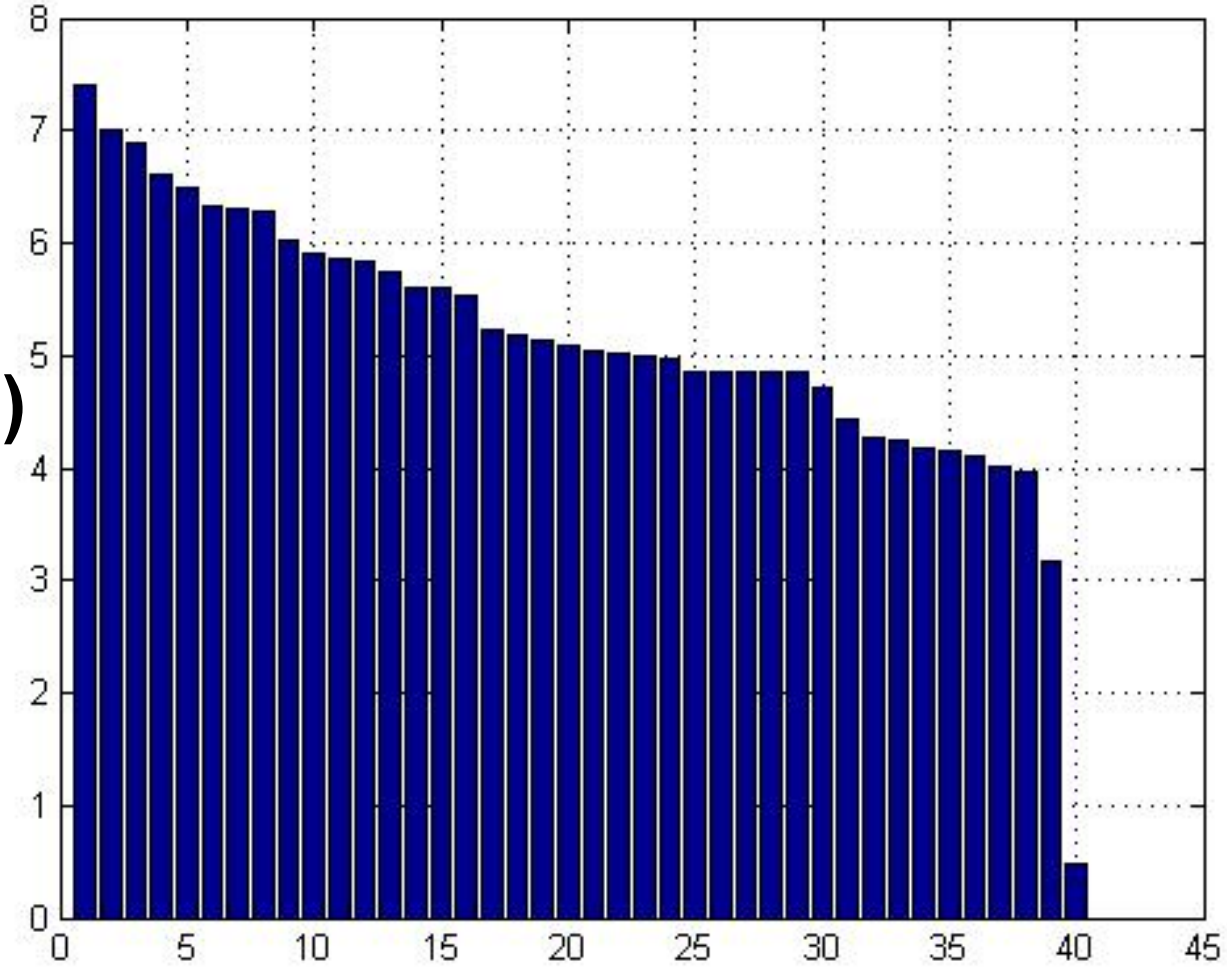


# „Koch“-Rezept nach (1) und (2) für 2013



# Koch“-Rezept nach (3), (4), (5)

**Log(Amplitude)**

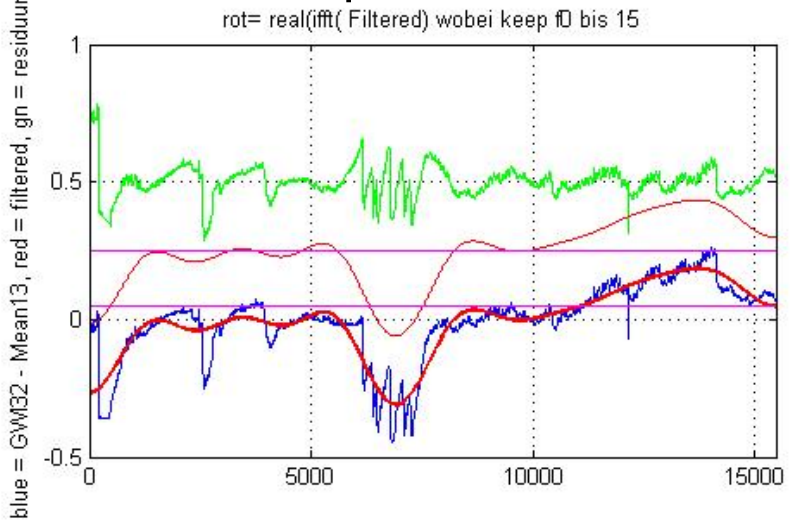



Sortiert,
Jahr
2013,
Gwl32

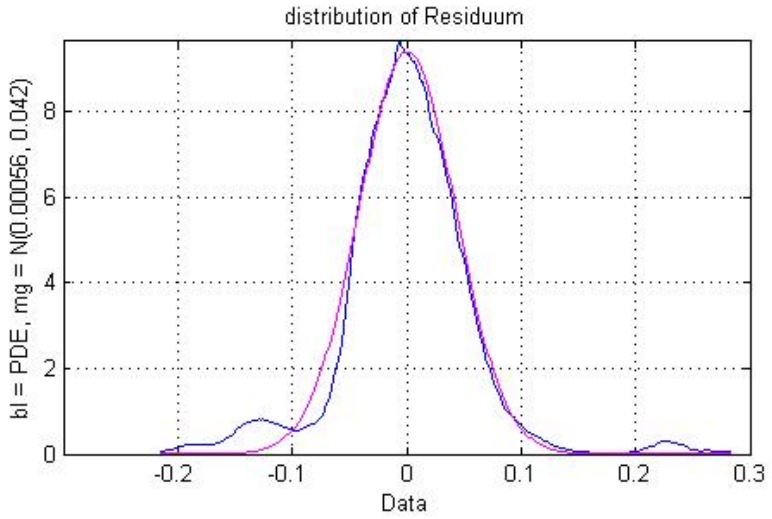
Anz. Frequenzen (Nyquist )

# **(6)** Tiefpassfilterung

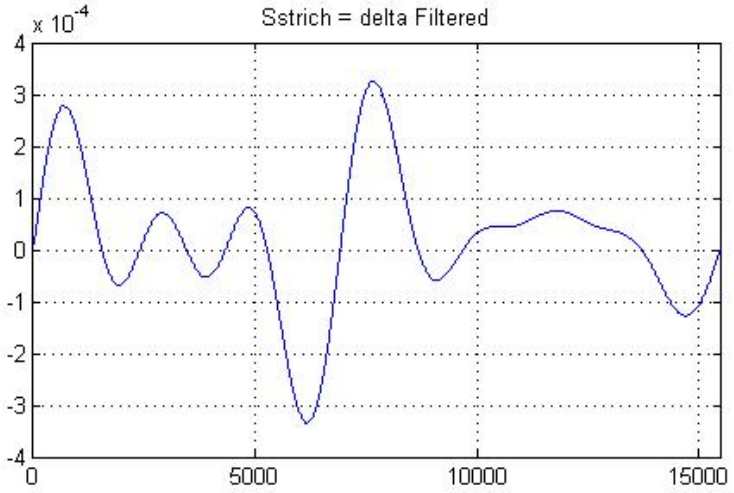
Die ersten 15 Frequenzen in rot, Residuum (grün)



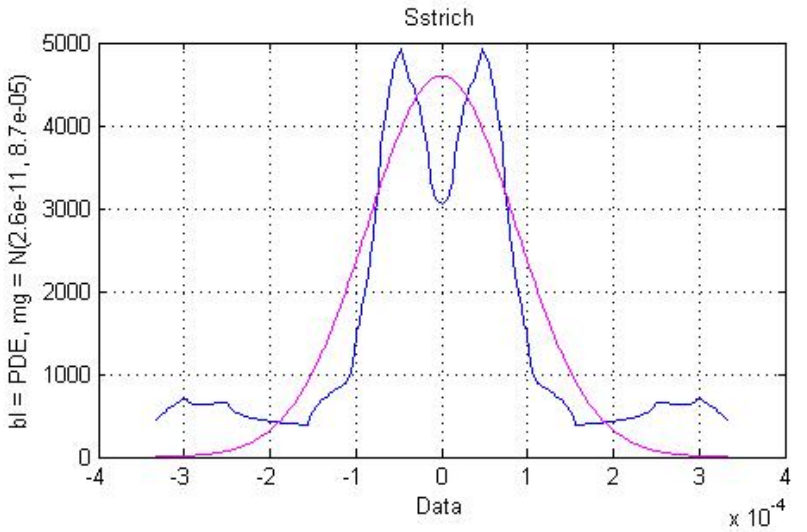

PDE Residuum




1.Ableitung der Season (oben in rot)




PDE der 1. Ableitung

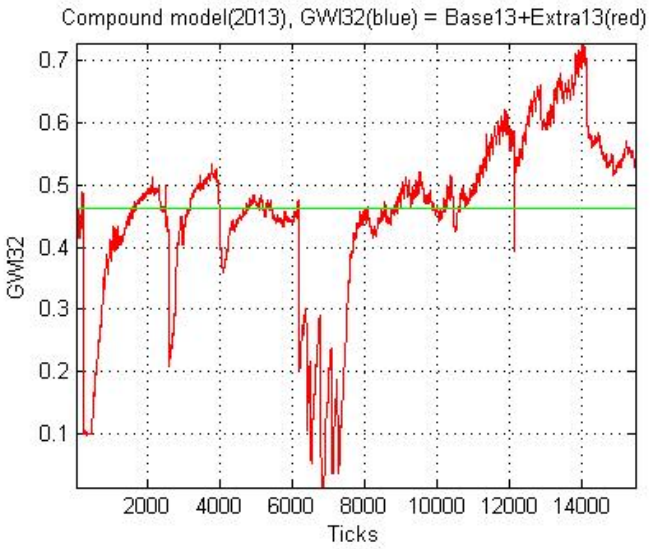


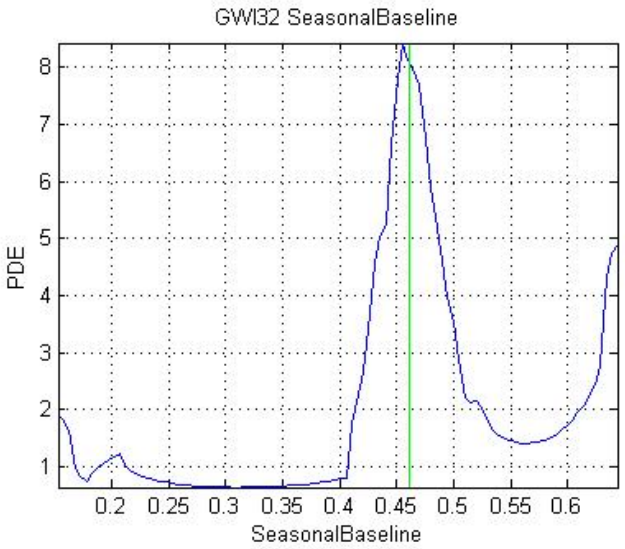
# Komponentenmodell 2013

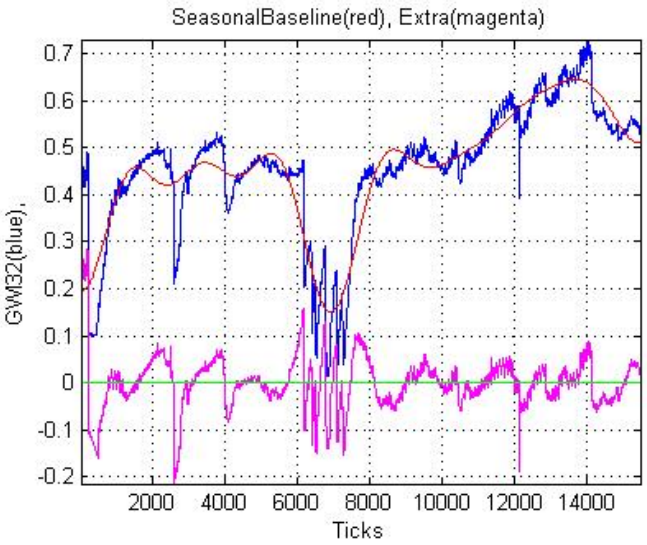
ZR: blau

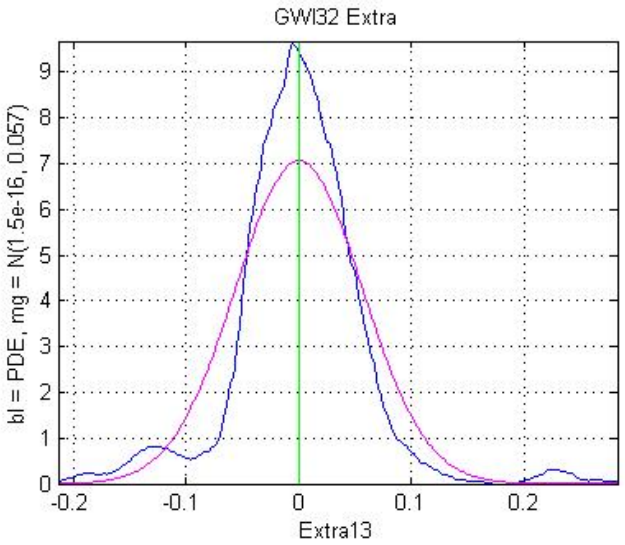
Season: rot

Residuum: magenta

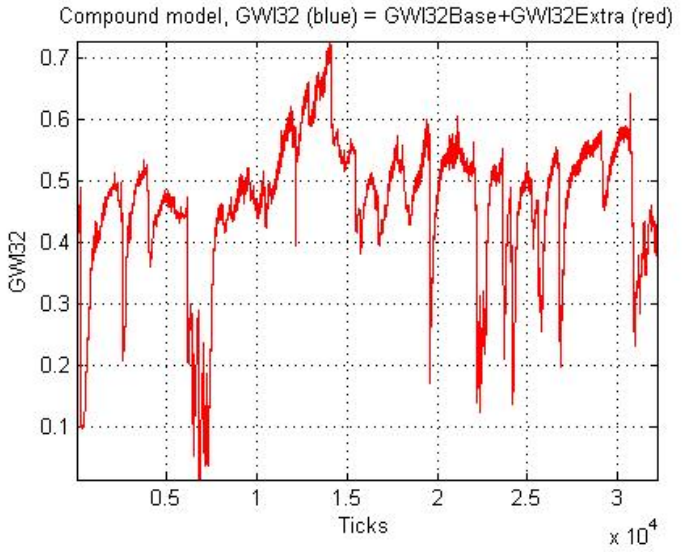


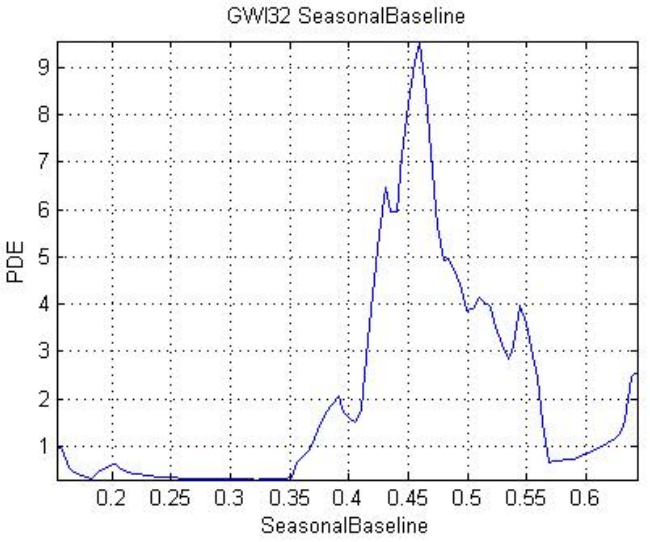



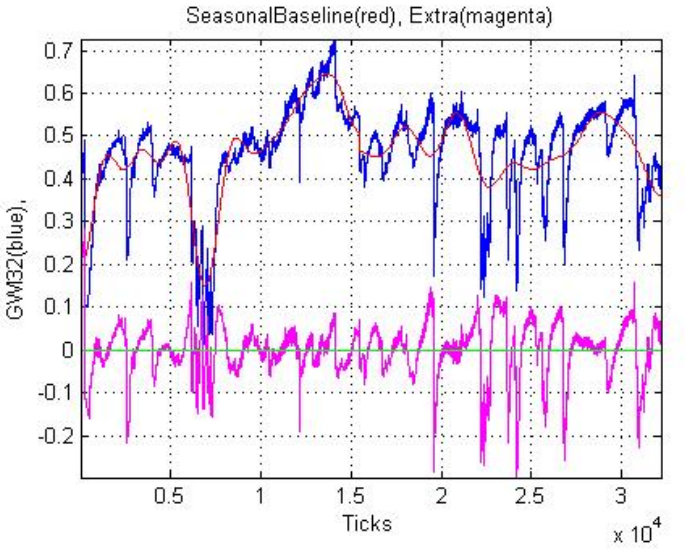



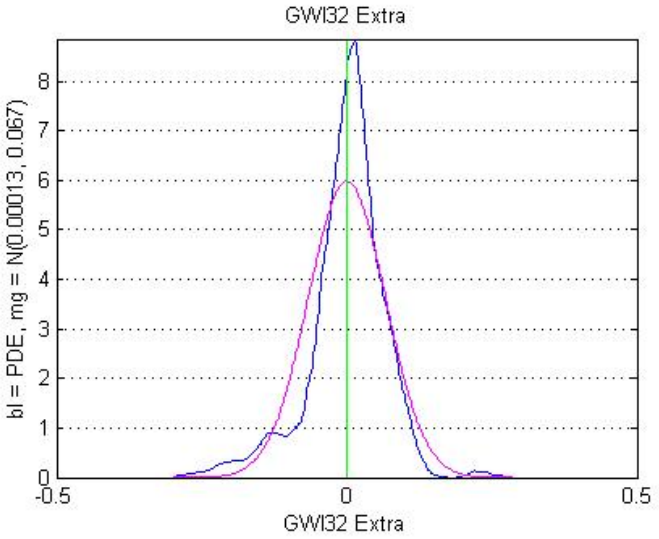



# „Koch“-Rezept (7): 2013 und 2014 GWI32



# After Preprocessing: Explain Nitrate

1.) Model: GMM of Nitrate-Extra

- Verification:
  - Number of Modes defined by AIC and BIC
    - GMM compared with EM-algorithm of different modes
  - GMM with QQplot, Kolmogorov-Smirnov-Test, Xi-Quadrat-Test

2.) Model: All variables were grouped according to Nitrate Modes

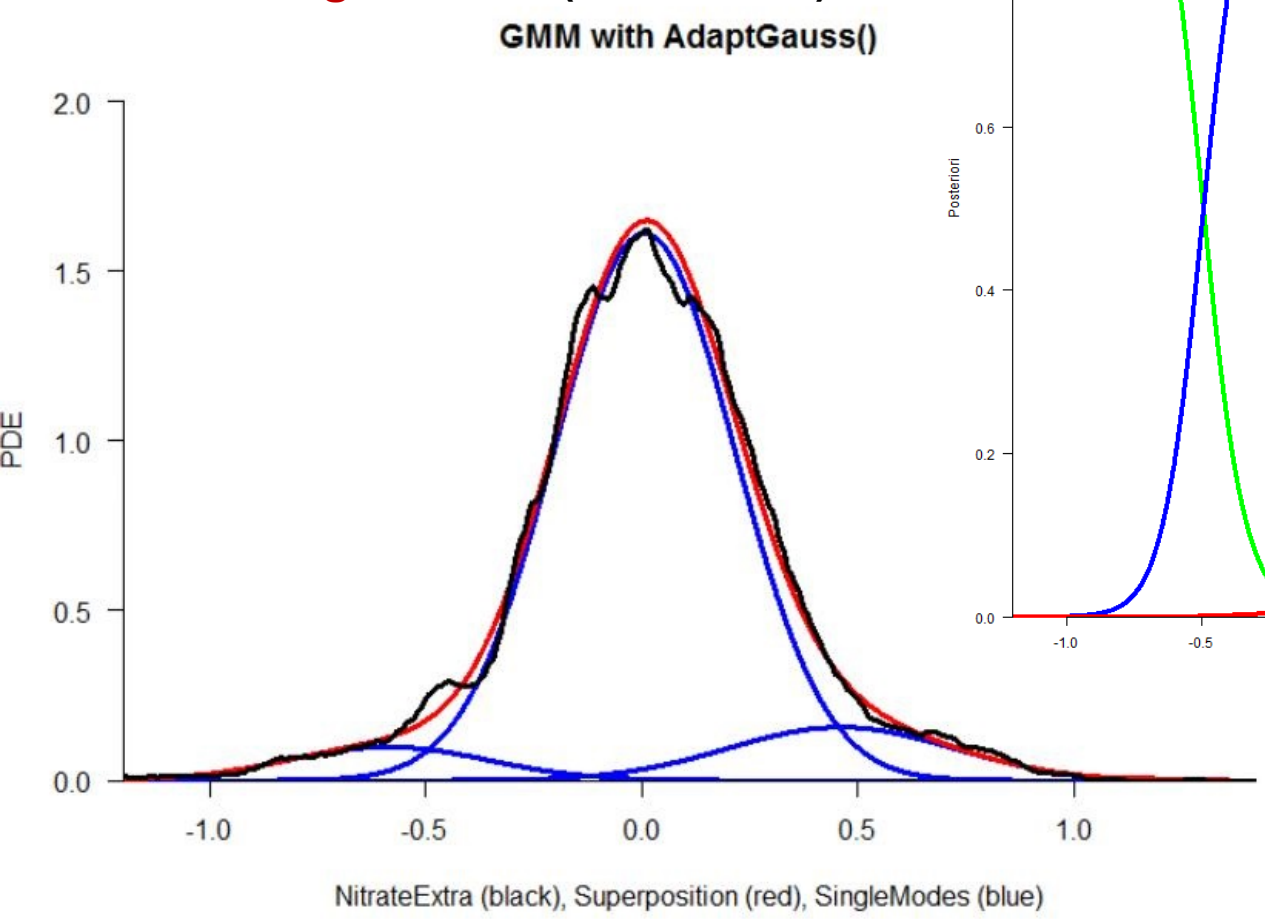
  - Interpretation because of: PDE
  - Verification: Bonferoni corrected two-sample t-test

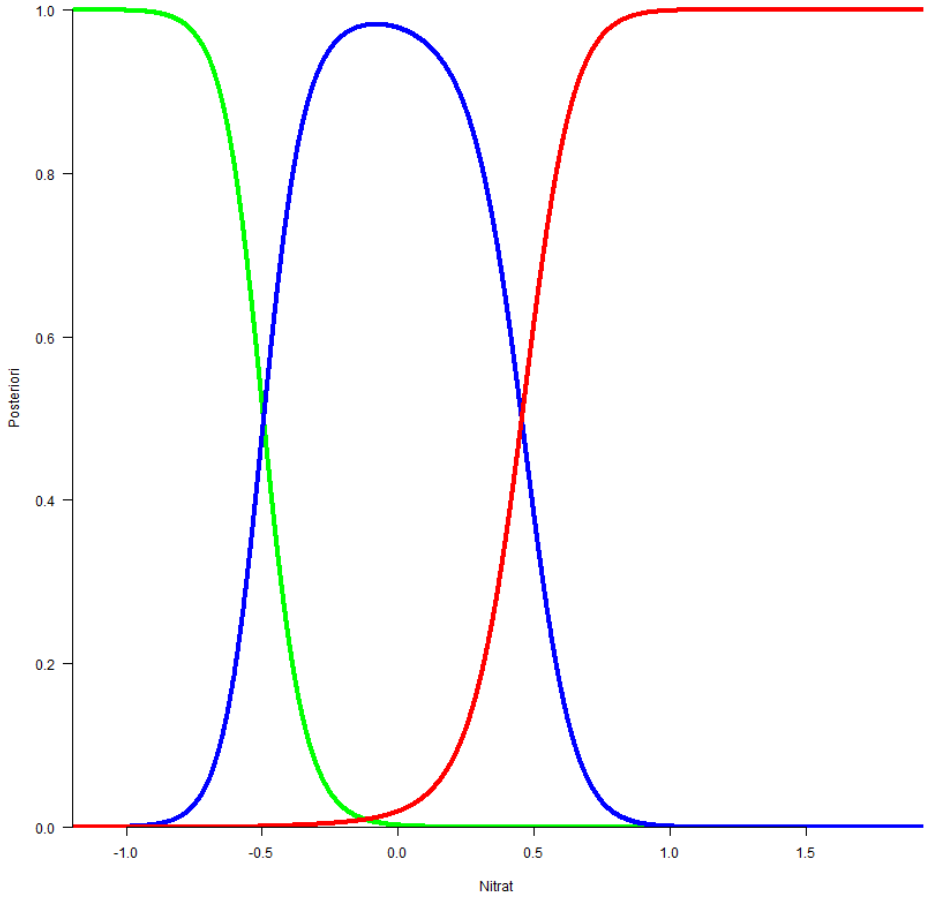
*-> Always verify your model*

*-> Verify using statistics and visually*

# GMM of Nitrate-Extra

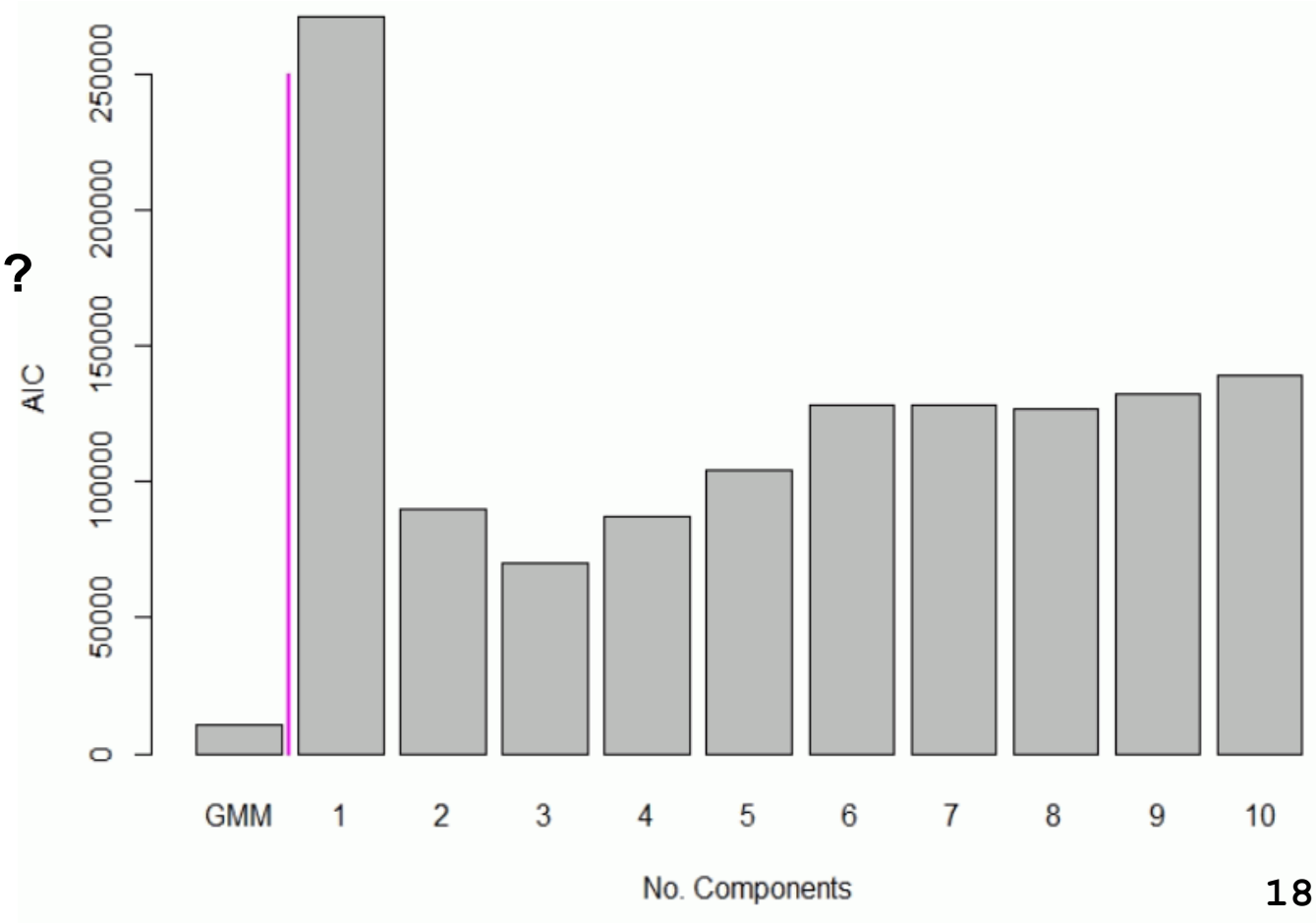
- **Explain semantically:**
  - **Low Nitrate (5% of data)**
  - **normal Nitrat (89% of data)**
  - **High Nitrate (6% of data)**



# Verification: Number of Modes

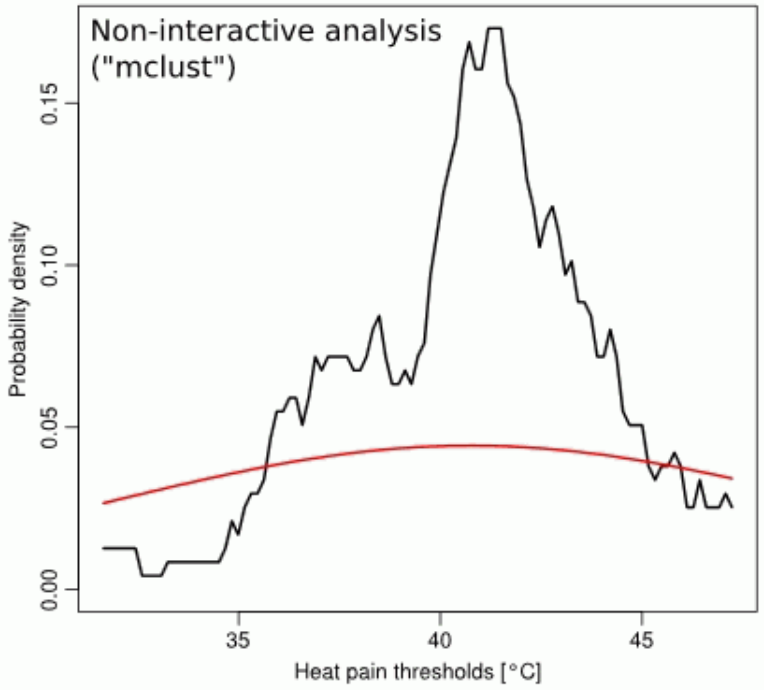
- **AIC by EM model from one to ten modes compared to 3 mode model**
  - **Akaike information criterion (AIC) measures quality of statistical models by maximum likelihood**
- **BIC analog**
- **Why is EM bad?**

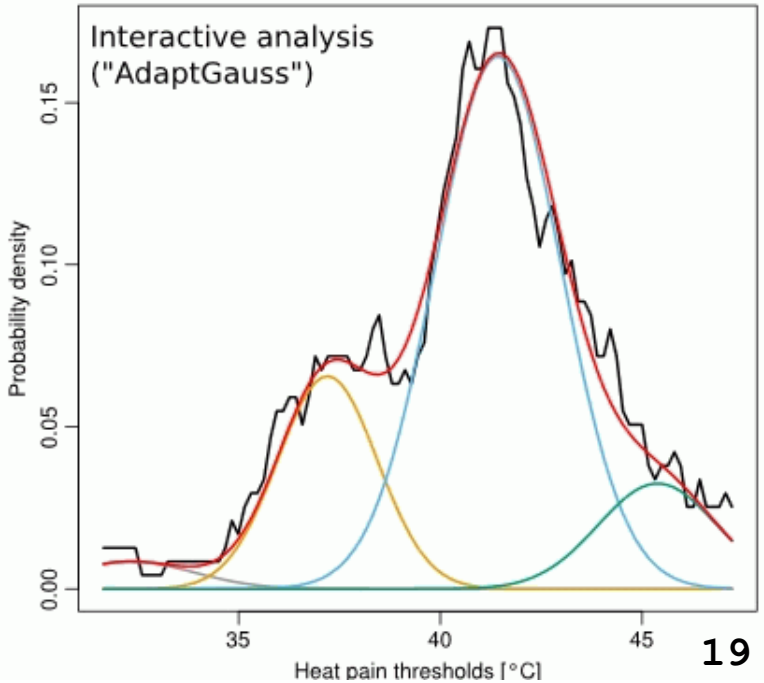


# EM-problems

1. K: number of modes
   - In example chosen by mclust R package
2. Starting values for Means, SDs, Weights
3. Optimizes „MaximumLikelihood“
   - Product!
   - Outliers overweighted
   - Good Optimization for borders of the distribution

Ultsch, A., Thrun, M.C., Hansen-Goos, O., Lötsch, J.: Identification of Molecular Fingerprints in Human Heat Pain Thresholds by Use of an Interactive Mixture Model R Toolbox(AdaptGauss), International Journal of Molecular Sciences, doi:10.3390/ijms161025897, 2015.



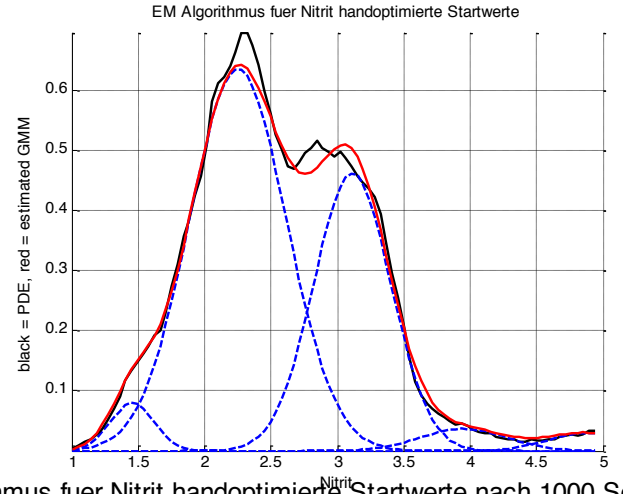
# Probleme mit EM: seltsame „Optimierung“

*Entommen aus der*

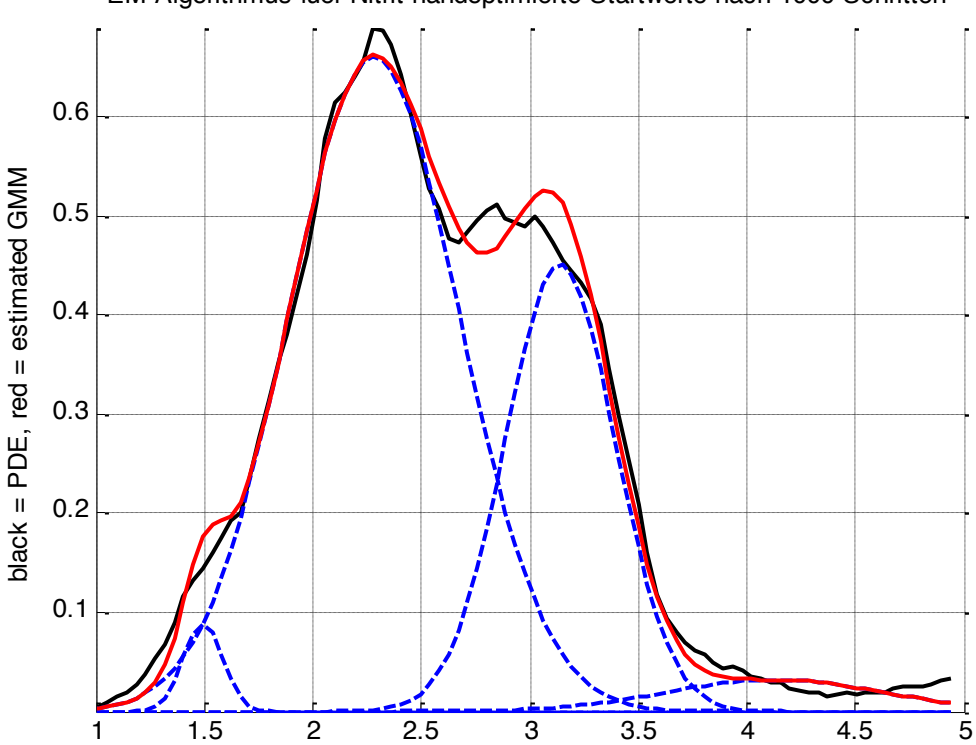
*Knowledge Discovery VL von*

*Prof. Dr. Ultsch*

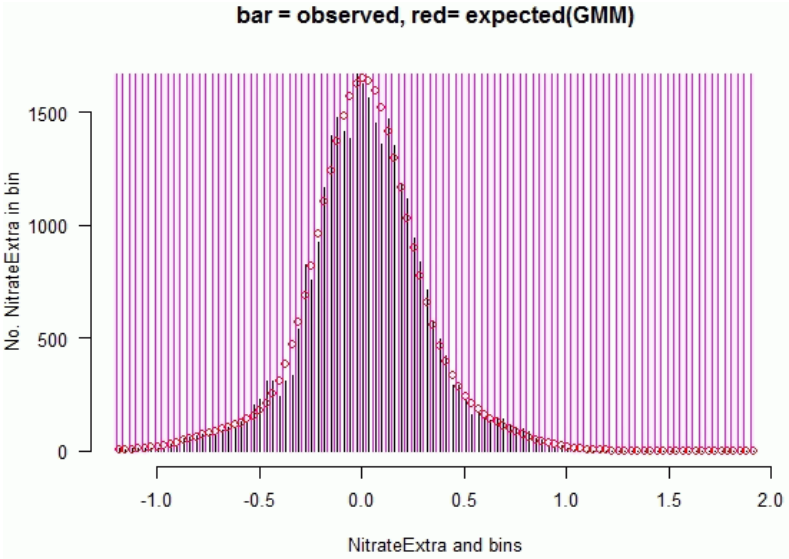
- **Startpunkt: händisch ausgewählt:**

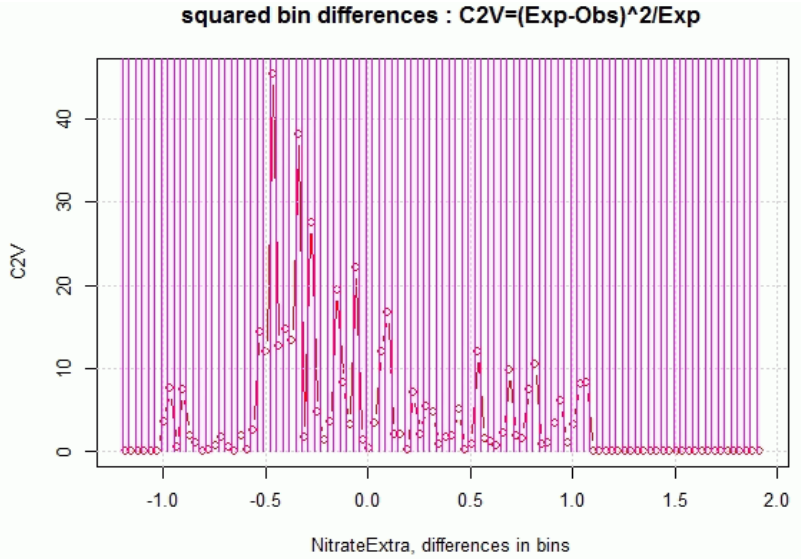


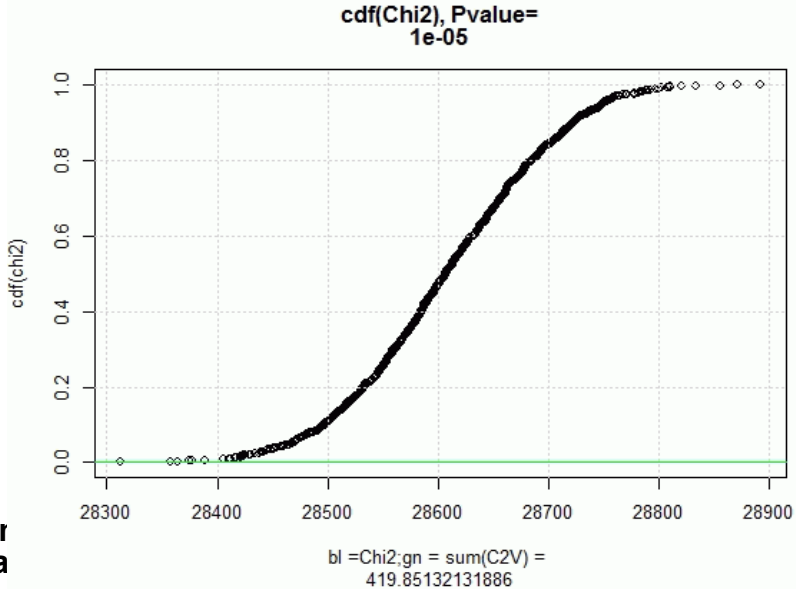

- **Nach 1000 EM Schritten: Moden semantisch nicht erklärbar**

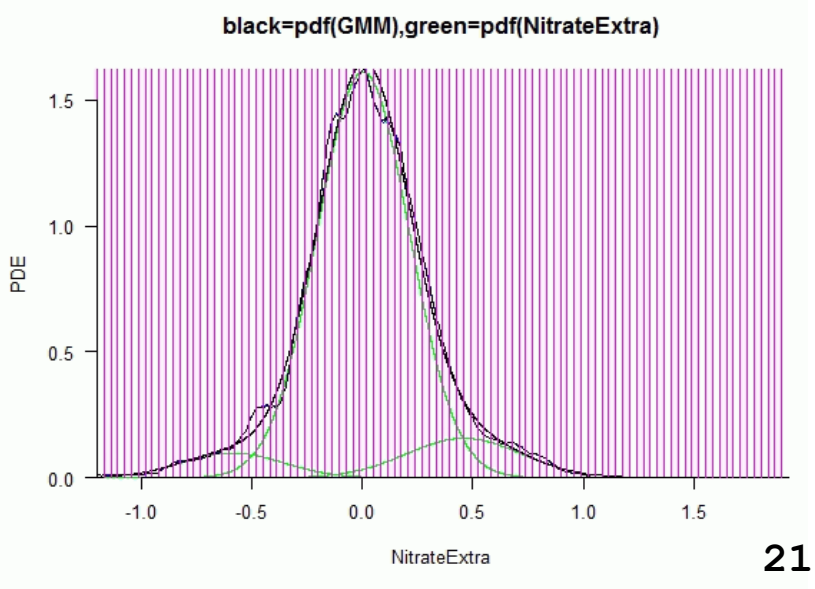


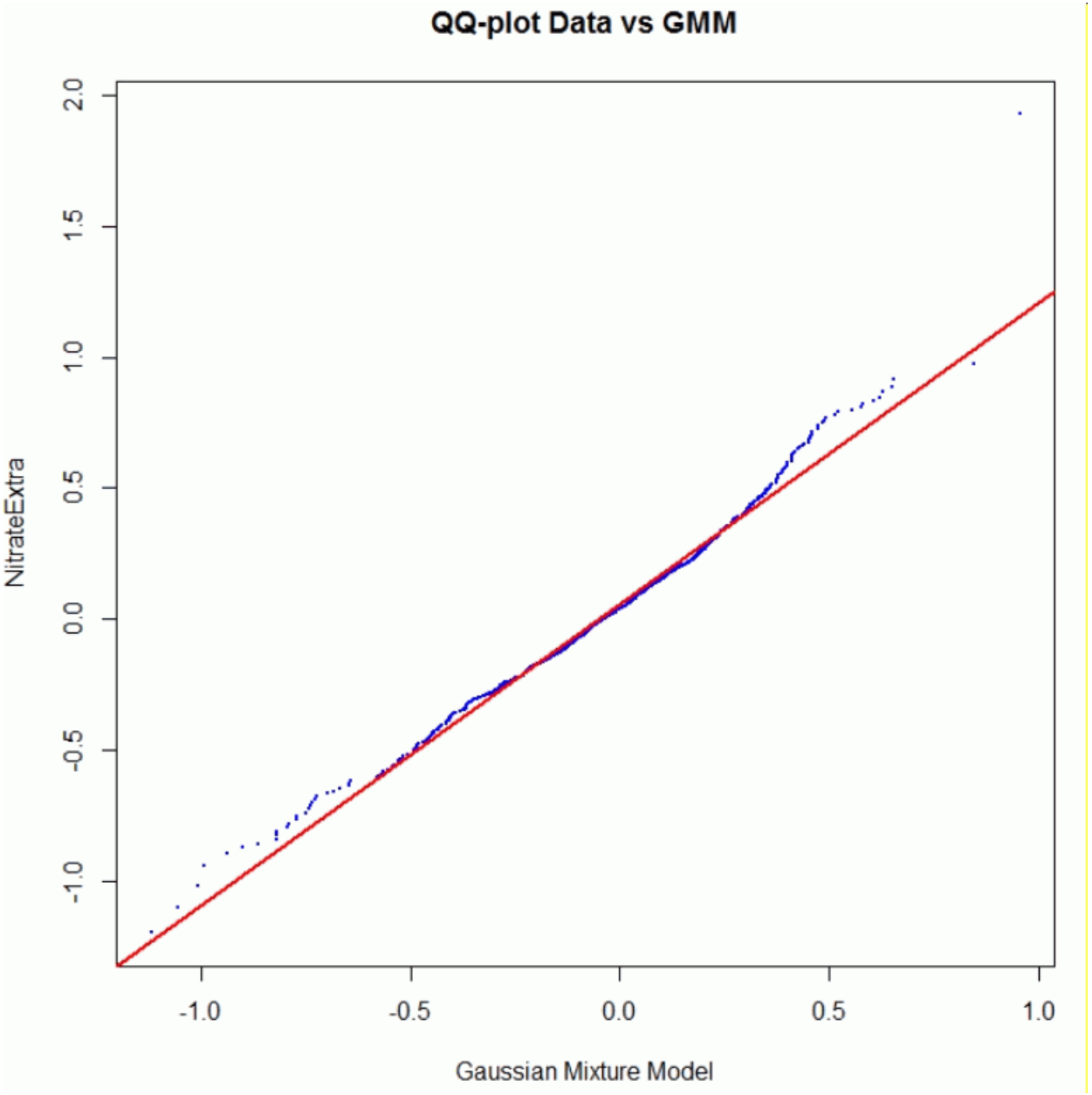
# Xi-Quadrat test



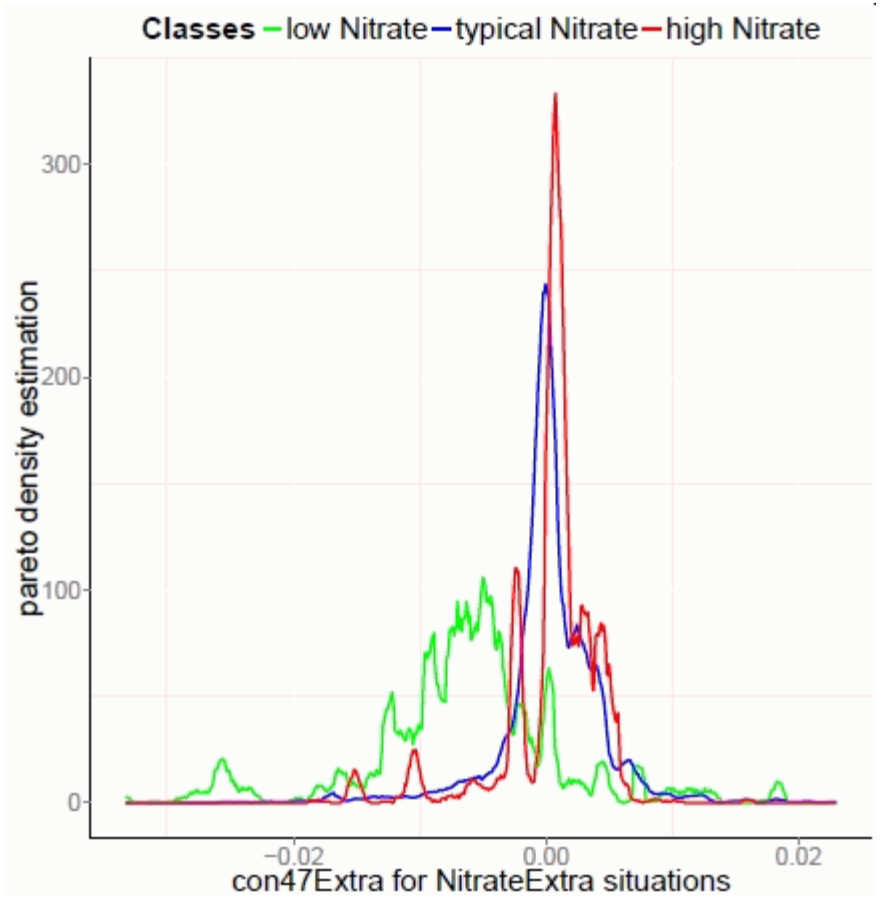



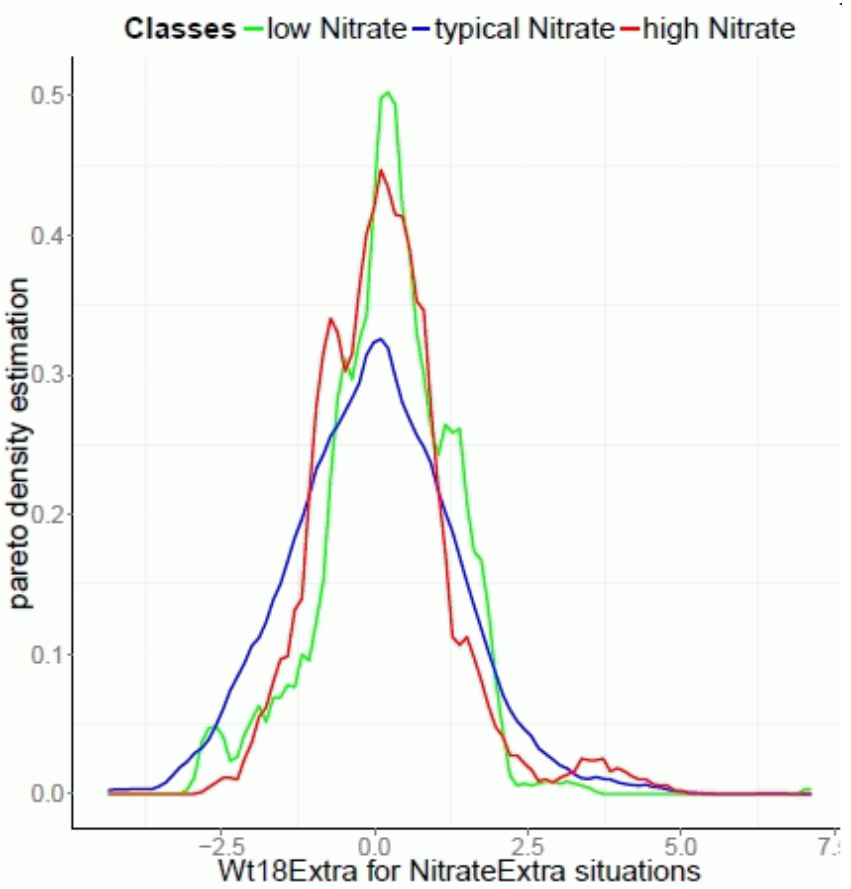




QQ-plot Data vs GMM
NitrateExtra
Gaussian Mixture Model
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5
1.0

# All variables were grouped according to Nitrate Modes

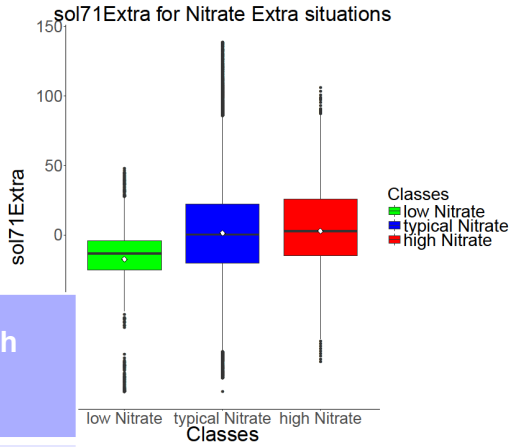
Visually: Significant versus non-significant

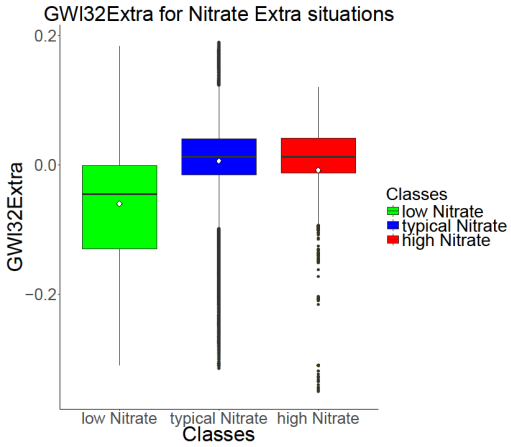


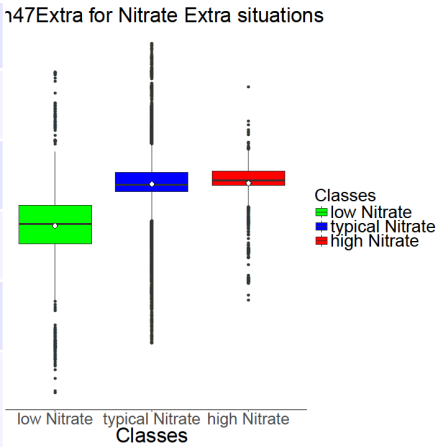



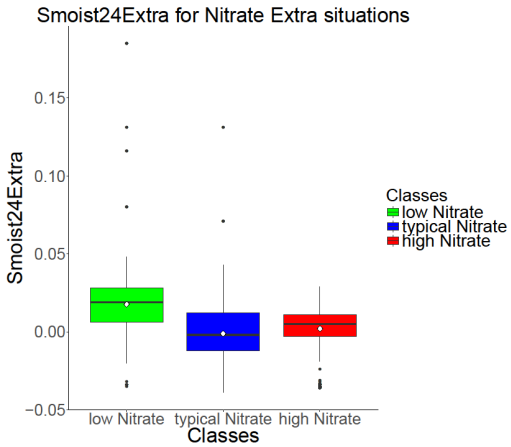
# All variables were grouped according to Nitrate Modes

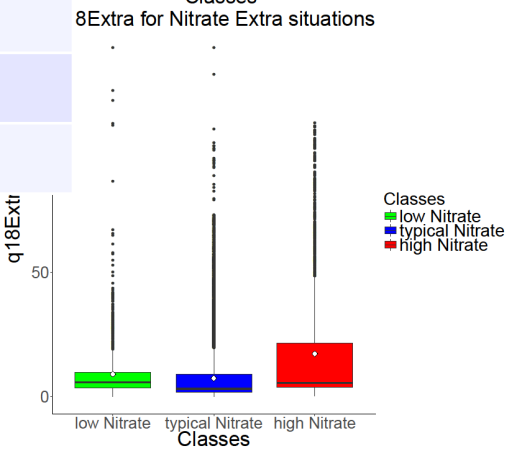
| Variable | low-typical | low-high | typical-high |
|---|---|---|---|
| GWlbg3-Extra | n.s. | 5.7e−10 | 2.9e−10 |
| GWlbg32-Extra | 1.9e-137 | 2.8e−57 | n.s. |
| Wt13-Extra | n.s. | 3.4e−09 | n.s. |
| Sol71-Extra | 5.1e−111 | 2.5e−81 | n.s. |
| Con47-Extra | 1.3e−205 | 6e−198 | n.s. |
| Smoist24-Extra | 3.5e−220 | 1.2e−125 | 1.5e−15 |
| q13-Extra | n.s. | 2.3e−31 | 3.2e−59 |
| q18-Extra | n.s. | 9.6e−40 | 6.9e−69 |

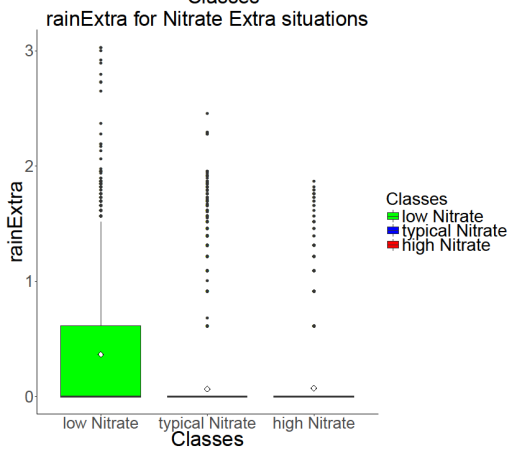



**Thrun - Databionics**
**University of Marburg**

# Results

- **Lowest Nitrate is driven by groundwater depth close to the surface and high soil moisture**

  **=> Indicates subsurface staturated state and hydrological connectivity**

  **=> wet conditions => Denitrification**

- **Highest Nitrate is dirven by high solar radiation and deeper groundwater but still moist soils**

  **=> Hydrological recession**

  **=> drying conditions => Nitrification**

# Analysing a data set

- Goal: Mostly given by domain/topical experts
  - Often lost in Translation!
- Preprocessing (here: compound model using FFT)
  - Time consuming, many trials and errors…
  - Good, if results are good…
- Analysis (here: GMM, PDE, statistics):
  - Verify your models statistically and visually!
- Interpretation by topical/domain experts
  - Important: Elaborate results!

# Thank you for listening, any questions?