# Teporales Data Mining, SS 2016

# Projekt 2, Wissen aus Hydrologie

# Gruppenmitglieder: Benjamin Rösner, Marta Lemanczyk, Jiachun Zhang

1. **Inspizieren sie die einzelnen Zeitreihen mit den Methoden der Knowledge Discovery**

Zur Untersuchung sind Datensätze „Chirimachay\_NO3.lrn“ , „Chirimachay\_Netradiation.lrn“ und „Chirimachay\_Precip.lrn“. Die Datensätze sind jeweils Liste mit 7 Variablen, so dass die Messungswerte und Messungszeiten genau notiert werden. Zuerst untersuchen wir den Datensatz „Chirimachay\_NO3.lrn“.

**Variable: Chirimachay\_NO3**

Die erste Erträge sowie die letzte Erträge sieht wie Folgende aus:

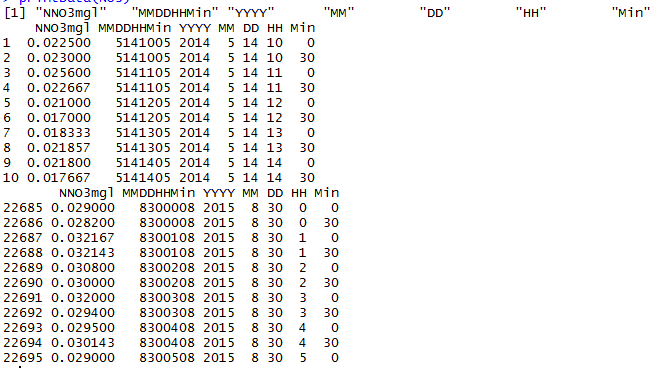


Figure 1: Überblick für den Datensatz „Chirimachay\_NO3.lrn“.

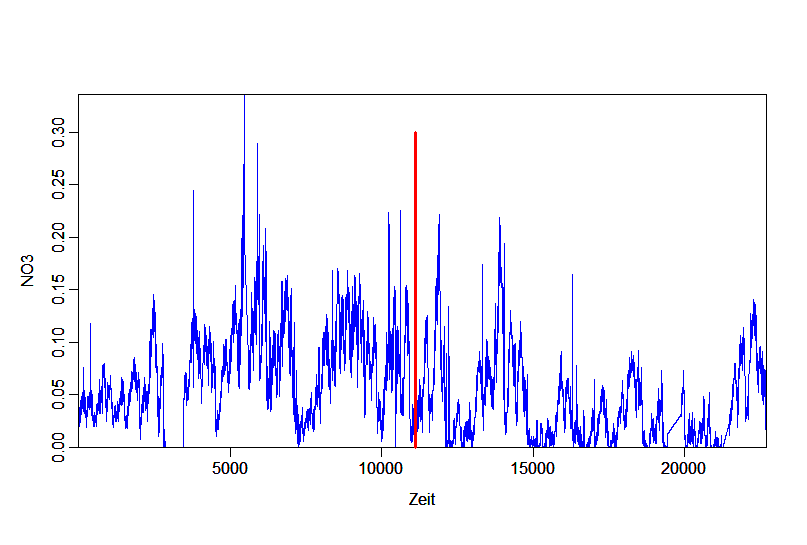


Figure 2: Überblick der Zeitreihe.

Es wird insgesamt 22695 Werte jede 30 Minuten von 10 Uhr, 14.5.2014 bis 5 Uhr, 30.8.2015 gemessen. Die Zeitreihe verläuft steitg, vorbei einige Fehlstellen durch Intepolation ersetzt werden kann, da es gibt Sprungen zwischen Zeitpunkten zu beobachten hat.

Die Zeitreihe ist nicht stationär, auch hetroskedastisch, daher ist die auch kein weißes Rauschen.

Für die Verteilung der zeitunabhängige Beobachtungswerte „NNO3mgl“ wirde folgende Histogramm, PDE plot, QQ-plot gegen Normalverteilung, Boxplot sowie das Barplot für NaNs ausgegeben.

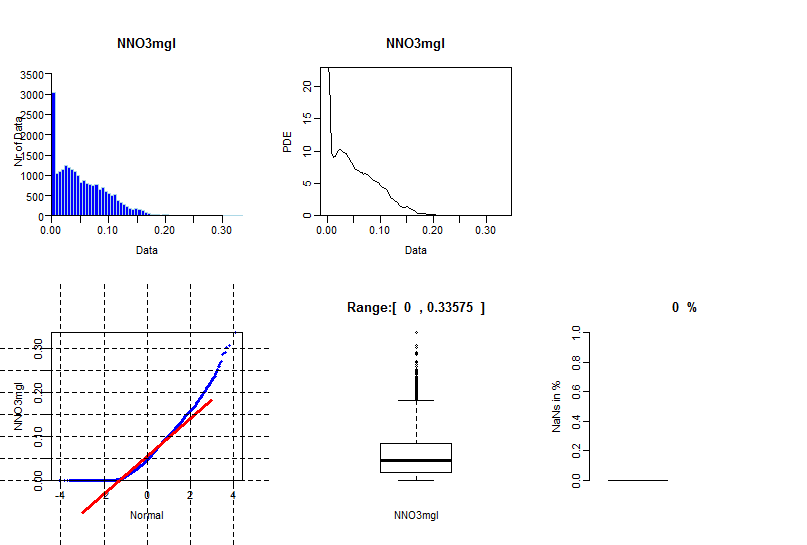


Figure 3: Verteilung von NNO3mgl.

* Keine NaNs, Es gibt ganz vielen Nullen. 🡪 NaNs durch Nullen eingesetzt.
* Wertbereich sehr klein.
* Nicht normalverteilt.

Versuche jetzt die extrem kleine Werte (< 0,05) wegzunehmen.

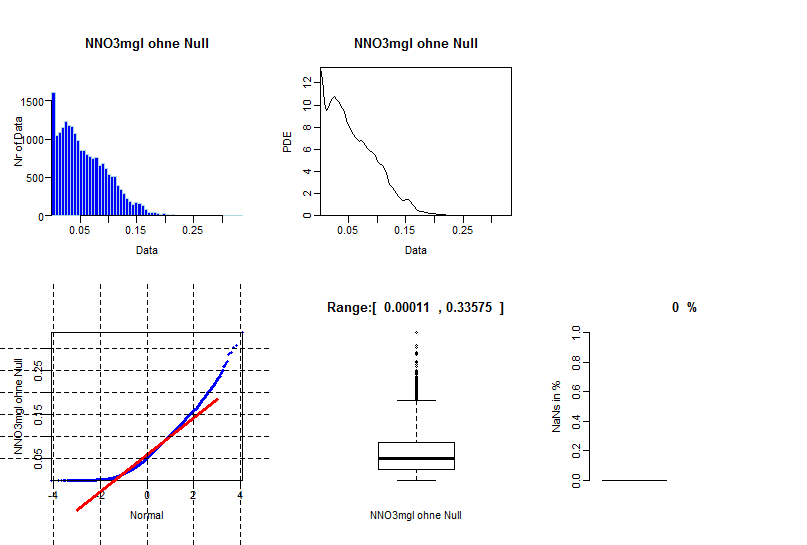


Figure 4: Verteilung von NNO3mgl, nach Eliminierung der Nullen.

* nach Eliminierung der Nullen ist es deutlich, dass es sich um einen Wachstum handelt. 🡪 logaritmieren?
* Wertbereich sehr klein 🡪 prozentuierung

Nach mehrfach Durfühung der Box-cox Transformation hat Faktor 0,5 als Empfehlung erhalten:

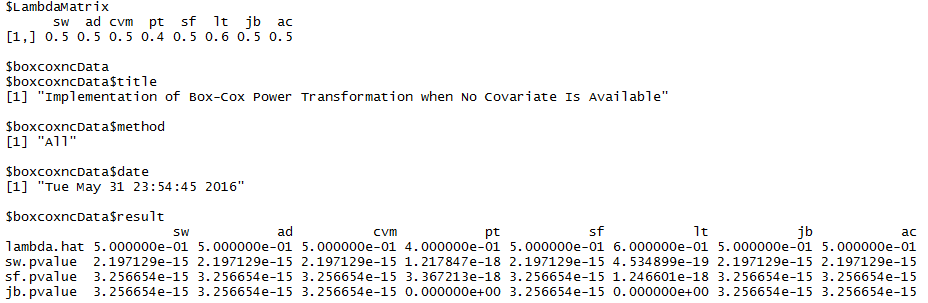


Figure 5: Eine der Ergebnissse bei Box-Cox Power Transformation

Betrachte daher die Transformation (Wurzel(Daten)\*100):

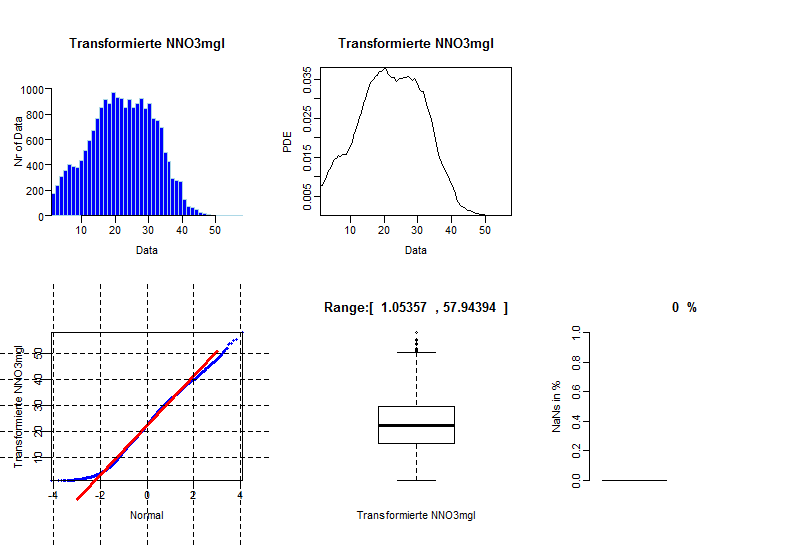


Figure 6: Transformierte Datensatz nach Box-cox Power Transformation.

* Die Approximation in der Mitte sehr gut, bei kleiner Werte ist Approximation sehr schlecht. Muss erneut angepasst werden.
* 🡪 Gaussian Mixture Modell

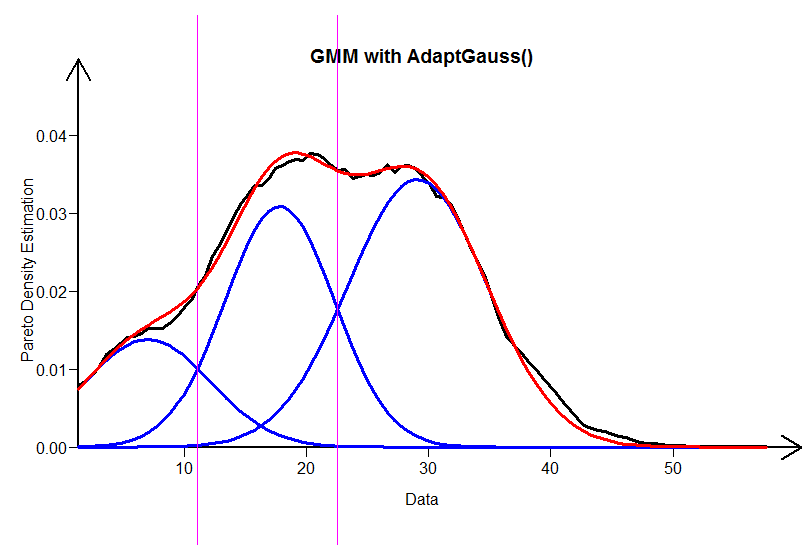


Figure 7: GMM bei Transformierte Daten, 2 Gauss.

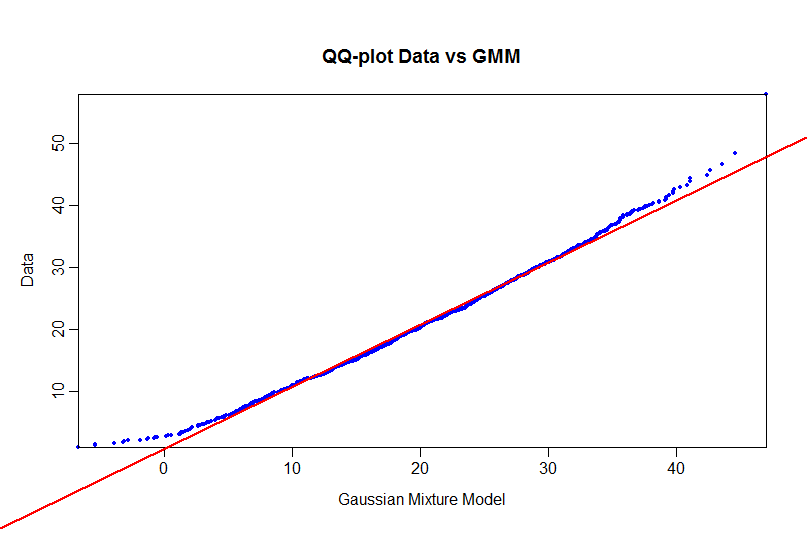


Figure 8: QQplot zur Verifizierung, Daten gegen GMM.

Hier sieht man, dass die Anpassung bei der extrem Werte Bereich immer noch schlecht ist.

Betrachte noch die statistische Verifizierung.

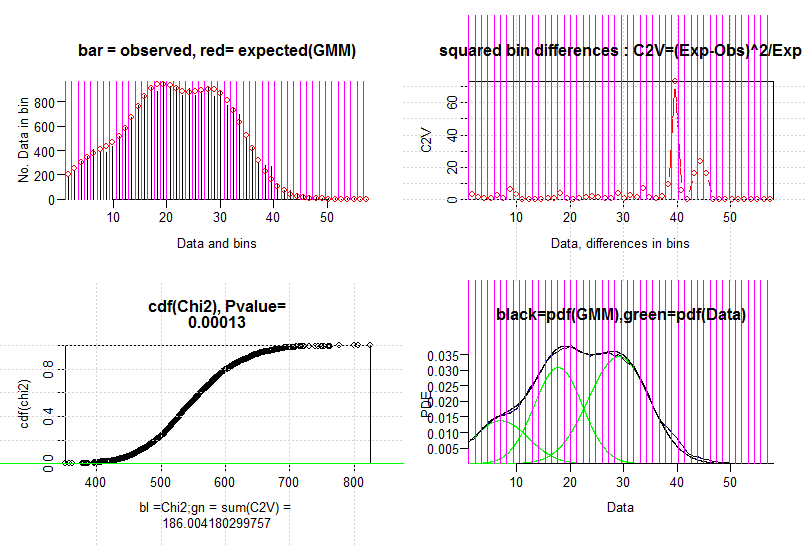


Figure 9: Chi-Quadrat-Test. H0: Daten und GMM besitzen die gleiche Verteilung.

P-Wert sehr klein. H0 wrid abgeleht. Anpassung schlecht.

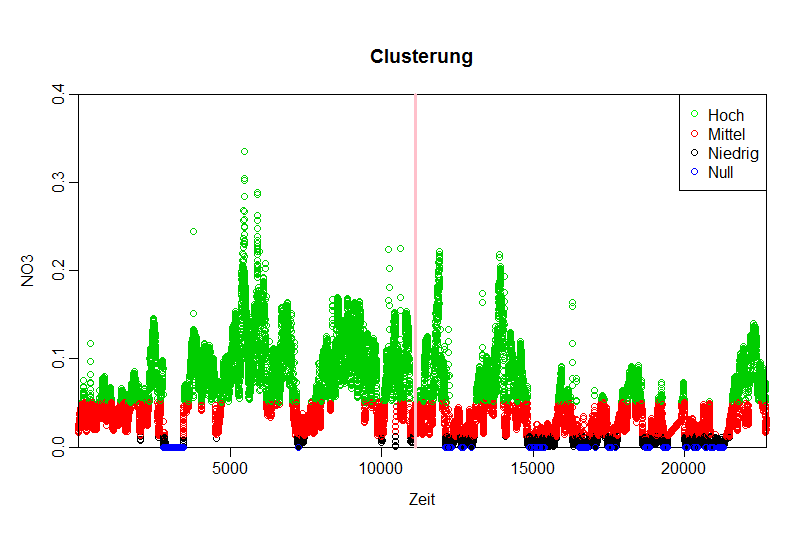


Figure 10: Clusterung nach GMM, extrem kleine Werte bilden sich als eine Gruppe.

Die Beobachtung kann zu viel Klassen geteilt werden, Niedrige Stoffkonzentration, mittele Stoffkonzentration hoch Stoffkonzentration sowie die Nullen.

**Variable: Chirimachay\_Netradiation**

**Variable: Chirimachay\_Precip**

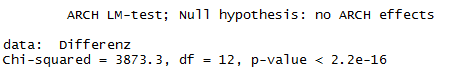
1. **Versuchen Sie jede der Variablen mit einem passenden Ansatz zu modellieren. Orientieren sie sich hierbei an der Vorlesungseinheit zur Hydrologie.**

**Variable NO3**

ARCH/GARCH

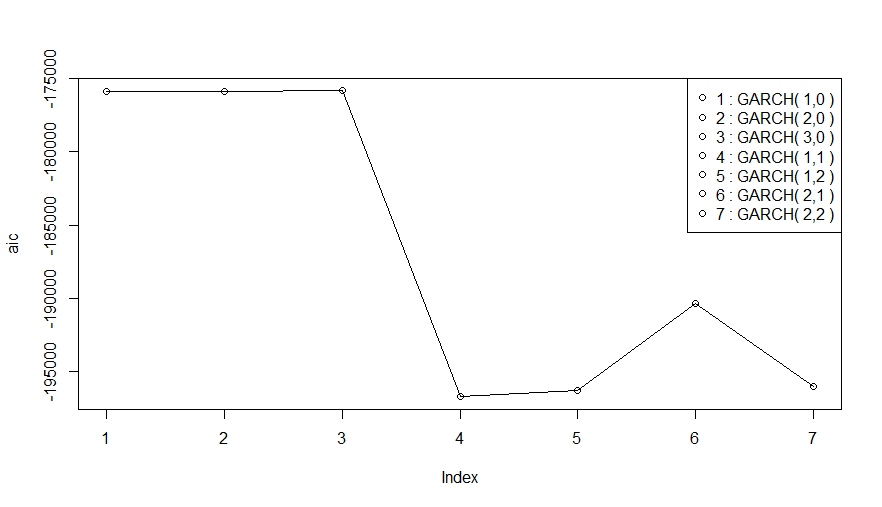
Differenzreihe betrachten.

ARCH LM-Test, teste, ob ARCH Effekte gibt:

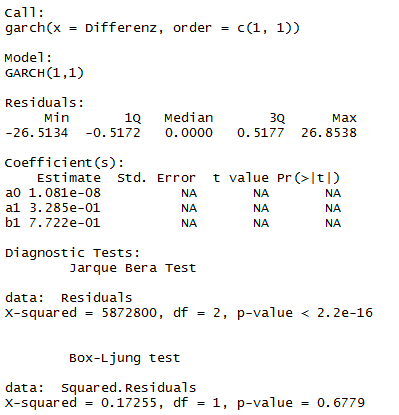


Nullhypothese wird abgelehnt, es gibt ARCH Effekt.

Die Parameter wird hier durch Anprobieren bestimmt, dann durch AIC Kriterium das beste Parameter bestimmt.

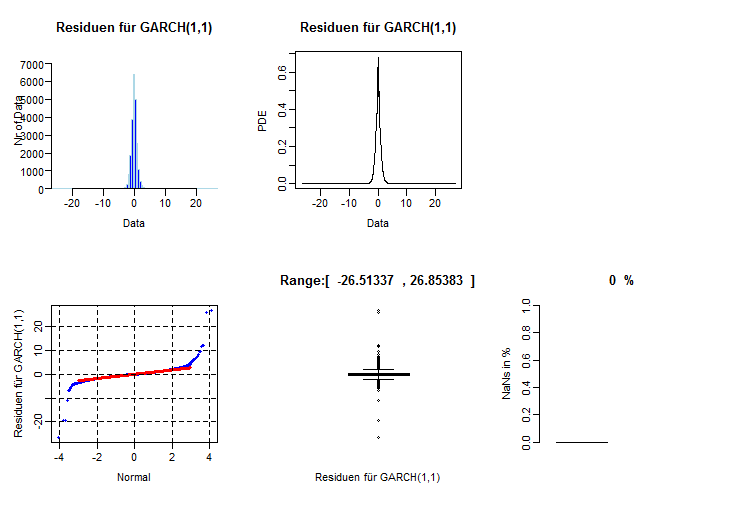


Also hier wird GARCH(1,1) ausgewählt.



Jarque Bera Test ist ein statistischer Test, der anhand der Schiefe und der Kurtosis in den Daten prüft, ob eine Normalverteilung vorliegt. Hier zeigt, dass die Residuen keine Normalverteilung ist.

Nach des Skripts muss Residuen Chi2 verteilt sein.



Die PDEplot sieht gut aus.

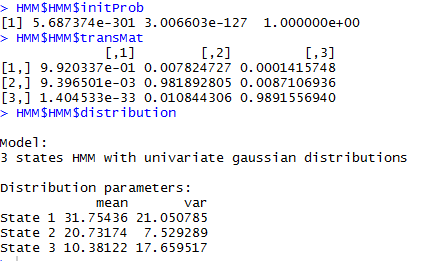
Box-Ljung test teste auf der Autokorelation, hier zeigt keine Autokorelation innerhalb der Residuen.

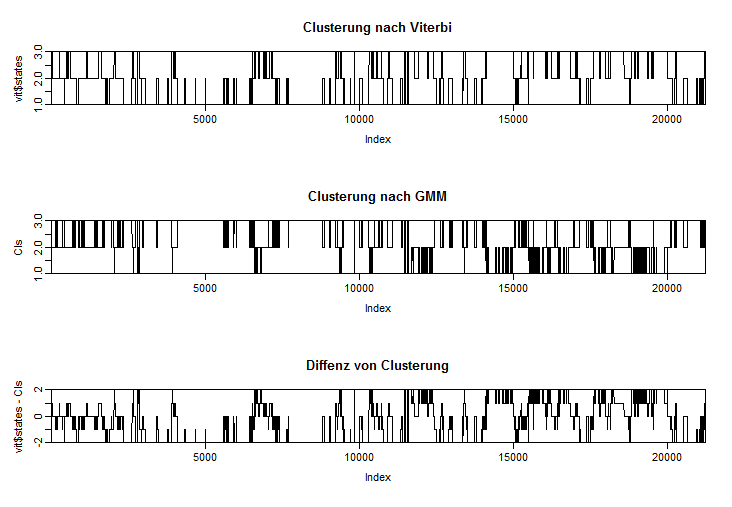
Nach des Skripts sollen Risiduen keine Autokorrelation vorliegen.

GARCH(1,1) ist eine gute Anpassung.

Hidden Markov Modell

NO3 ohne Nullen hat 3 Gaussian Mixture, s.o, nutze daher Transformierte Daten und 3 Bestände.





HMM liefert ganz anderes Ergebnis als GMM.

1. **Beschreiben Sie den Wissensgewinn durch diese Modelle. Wie und wodurch wird NO3 beeinflusst?**
2. **Modellieren Sie die halbstündlichen Änderung des NO3 mit AR, MA oder ARMA, begründen Sie die Auswahl ihres Modells. Finden Sie zuerst ein geeignetes R package dafür.**