Karlsruher Institut für Technologie

Geodatenanalyse I: Bivariate Statistik

Kathrin Menberg

Stundenplan

	08:30 – 12:30 Uhr	13:30 – 17:30 Uhr
Montag	Tag 1 / Block 1	Tag 1 / Block 2
Dienstag	Tag 2 / Block 1	**Tag 2 / Block 2**
Mittwoch	Tag 3 / Block 1	Tag 3 / Block 2
Donnerstag	Tag 4 / Block 1	Tag 4 / Block 2
Freitag	Tag 5 / Block 1	Tag 5 / Block 2

- **2.4 Bivariate Statistik**
- 2.5 Multivariate Statistik
- 2.6 Zeitreihenanalyse

Lernziele Block 2.4

Am Ende der Stunde werden die Teilnehmer:

- … mit den Konzepten von Kovarianz, Randverteilungen und Copulas vertraut sein.
- … verschiedene Korrelationskoeffizienten kennen und diese differenziert auf Geodaten anwenden können.

Bivariate Statistik

- Meistens wird an Standorten mehr als nur eine Messgröße aufgenommen
- Bivariate Statistik untersucht die Art und Stärke der Beziehung zwischen **zwei** Messgrößen
- Nomenklatur:
  - x: unabhängige Variable
  - y: abhängige Variable

O_2
Temperatur

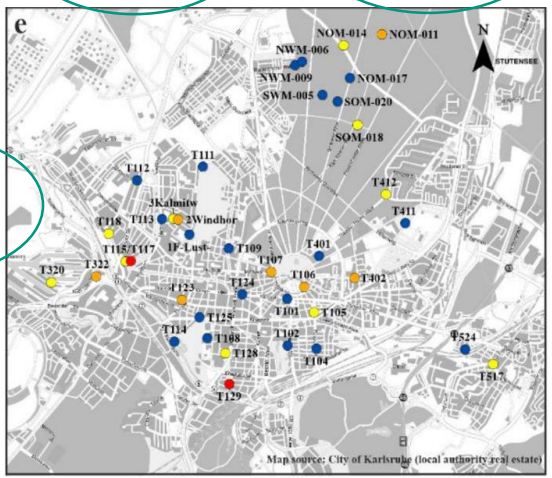
Eisen
Mangan

pH

Nitrat
Phosphat

Tiefe

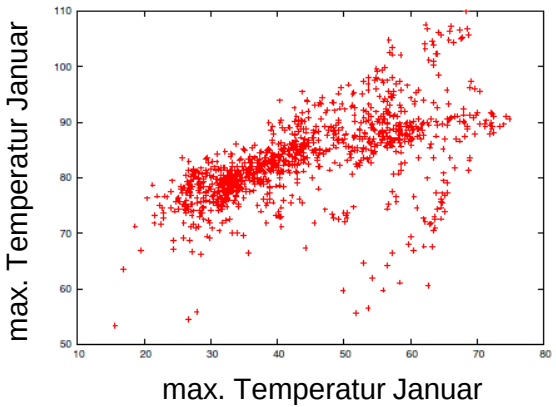
usw.

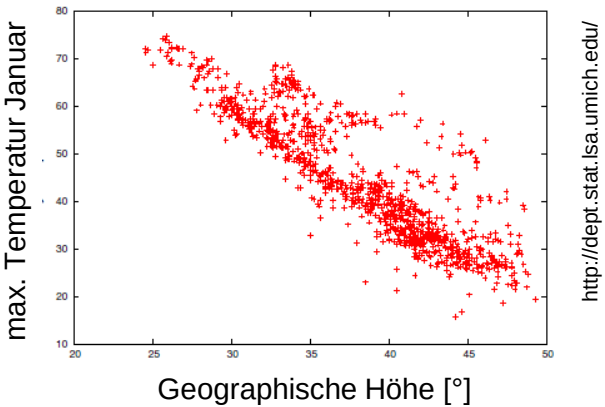


Koch et al. (2020)

Scatterplots

- Wichtigste graphische Zusammenfassung von bivariaten Daten

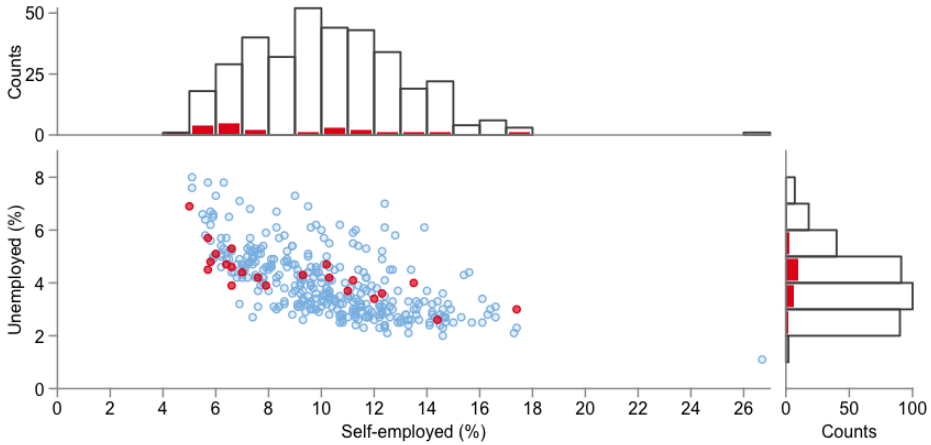




- Positive oder negativer Zusammenhang → Trend
- Keine Aussage über kausalen Zusammenhang möglich!

Randverteilungen

- Bivariate Datensätze haben gemeinsame (joint) und Randverteilungen (marginal distribution)

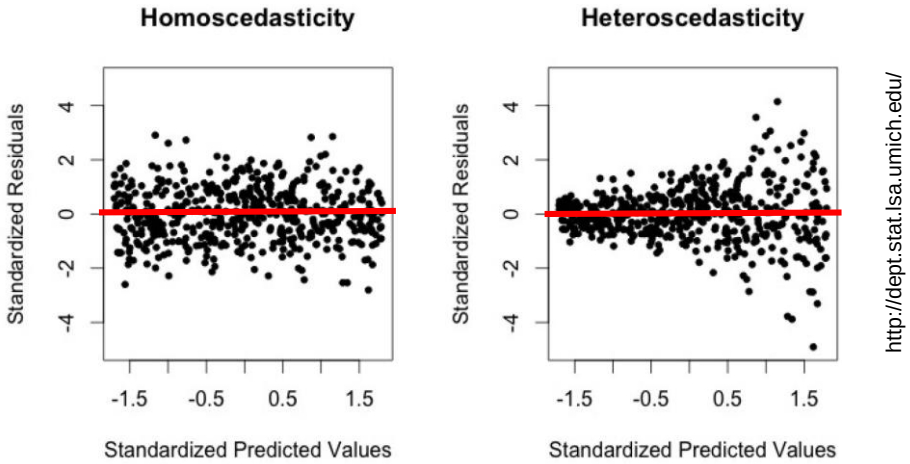


http://dept.stat.lsa.umich.edu/

- Das gleiche gilt für statistische Parameter, usw.

Trend ist nicht alles

- Trend: Zusammenhang der Mittelwerte



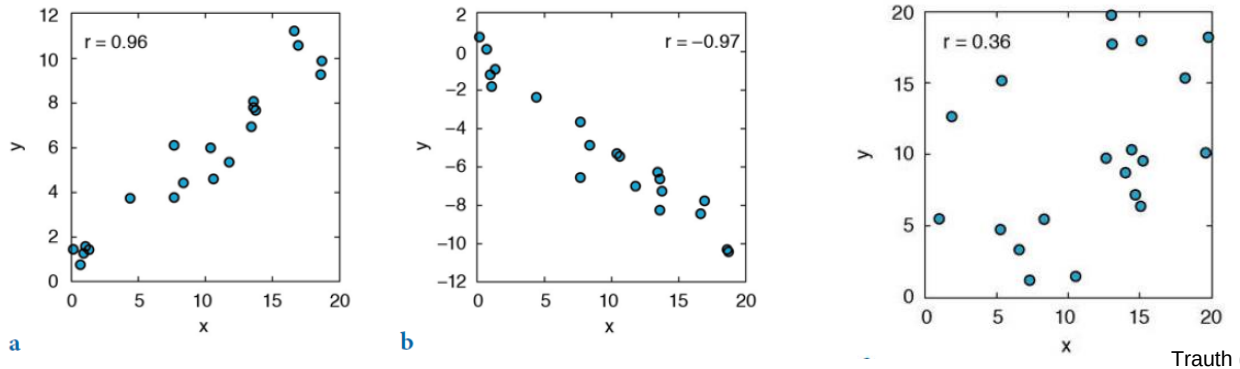
- Heterogenität der Varianz: Heteroskedastizität (heteroscedasticity)

Varianz und Kovarianz

- Varianz (σ^2): Streuung eines Parameters
- Kovarianz (covariance): $cov_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$
- Maß für die Assoziation zwischen zwei Zufallsvariablen
  - $cov_{xy} > 0$: hohe (niedrige) Werte von *x*, gehen mit hohen (niedrigen) Werten von *y* einher
  - $cov_{xy} < 0$: hohe (niedrige) Werte von *x*, gehen mit niedrigen (hohen) Werten von *y* einher
  - $cov_{xy} = 0$: kein Zusammenhang zwischen *x* und *y*

Korrelationskoeffizienten

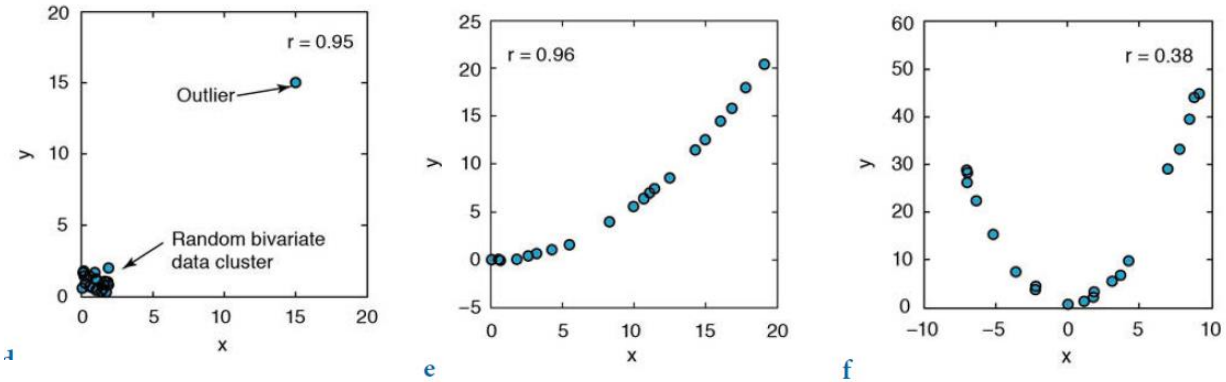
- Pearsons Korrelationskoeffizient $$\rho = \frac{Kovarianz}{std\ (x)\ \cdot std\ (y)}$$
- Maß für die Stärke des linearen Trends von (*x*, *y*)



Trauth (2015)

Korrelationskoeffizienten

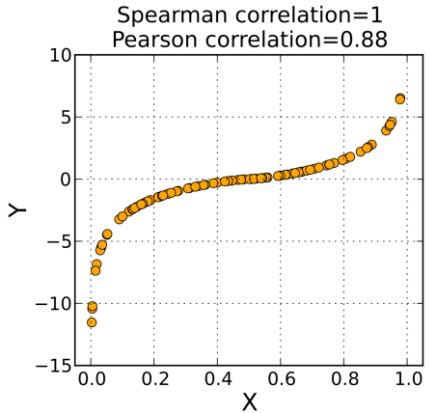
- Ausreißer und nicht-lineare Verbundenheit



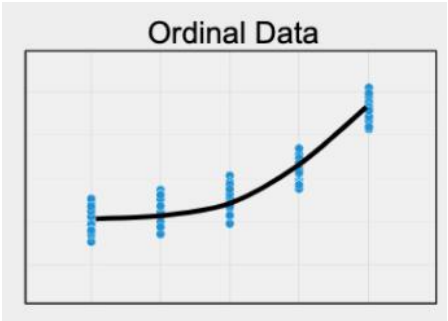
Trauth (2015)

Rang-Korrelationskoeffizienten

- Spearmans rho (ρ)
  - monotone Funktion
  - Statt Werte, Rang (ranking) der Daten



- Kendalls tau (τ)
  - Ähnlichkeit der Ränge von x und y
  - Robust gegenüber Ausreißern
  - Auch für ordinale Daten geeignet



www.wikipedia.org

Korrelationskoeffizienten

- Interpretation von Korrelationskoeffizienten:
  - Beschreibung der Stärke der Verbundenheit (association)

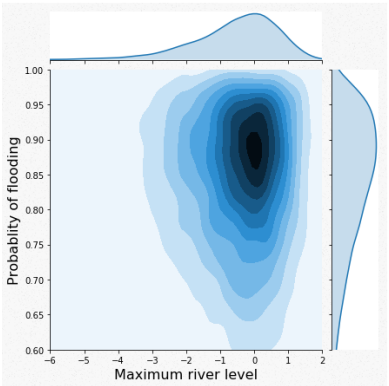
Perfekte Korrelation	ρ = ± 1
Hohe Korrelation	± 0.5 < ρ < ± 0.99
Moderate Korrelation	± 0.3 < ρ < ± 0.49
Niedrige Korrelation	0 < ρ < ± 0.29
Keine Korrelation	ρ = 0

- Abhängige Formulation der Messgrößen wenn $-\frac{2}{\sqrt{n}} > \rho > \frac{2}{\sqrt{n}}$
- Aussagekraft von Korrelationskoeffizienten mit *p*-Wert angeben.

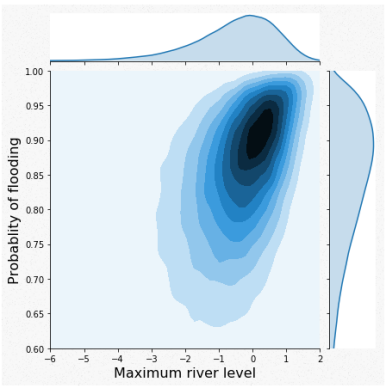
Copulas

- Randverteilungen von Zufallsvariablen die korrelieren
- Copula = „coupling function“ zwischen gemeinsamer Wahrscheinlichkeitsverteilung und Randverteilungen
- Erzeugen von zufälligen Wertepaaren mit Korrelation

ohne Korrelation



mit Korrelation

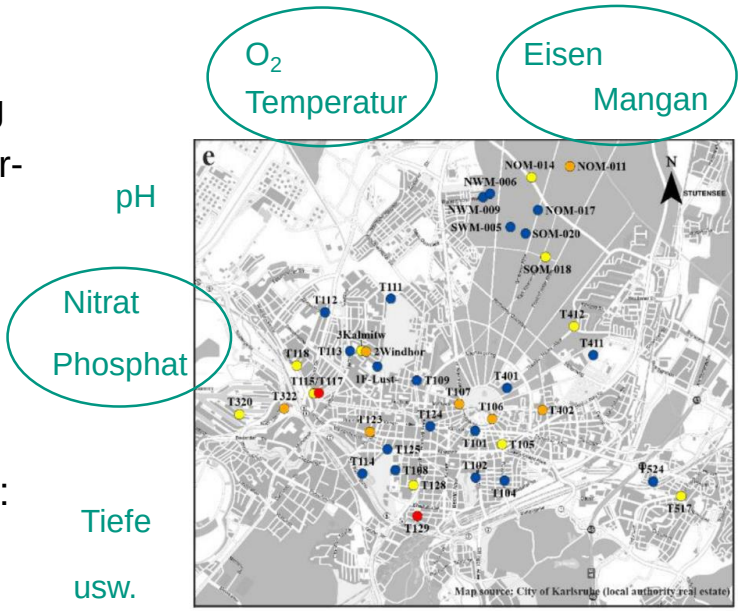


www.twiecke.io

Übung 2.4: Bivariate Statistik

- Grundwasserdatensatz Karlsruhe
  - Graphische Darstellung
  - Quantifizierung der Beziehung zwischen einzelnen Parameter-Paaren
  - Kovarianzen
  - Korrelationskoeffizienten

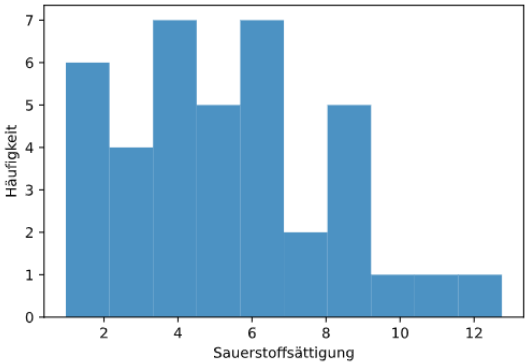
- Aufgaben in Jupyter Notebook: geodatenanalyse_1-2-4

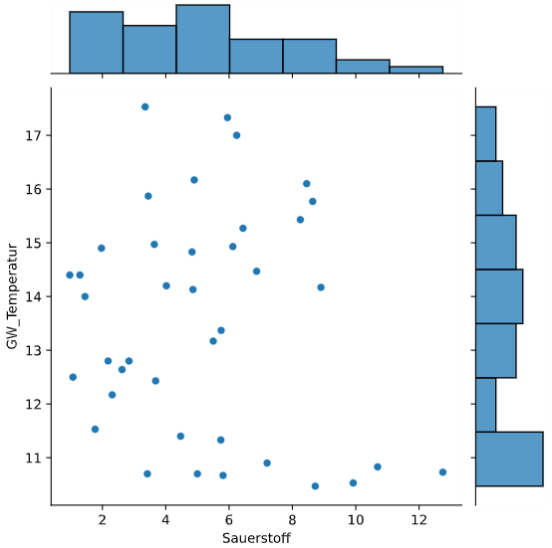


Koch et al. (2020)

Aufgabenbesprechung

- Visualisierung mit seaborn





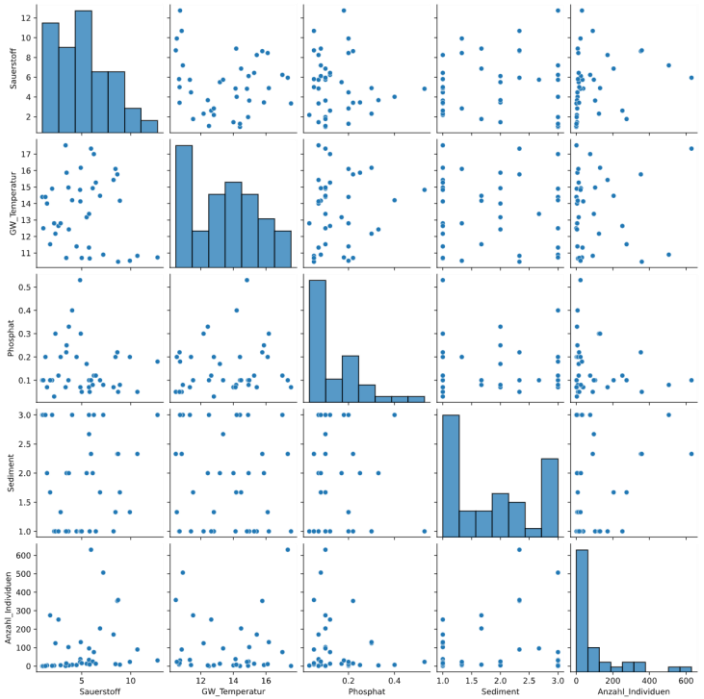
Aufgabenbesprechung

- Kovarianz (P_GWT) = 0.03
- Alle Kovarianzen:

```
cov_matrix = data.cov()
print(cov_matrix)
```

```
                  Sauerstoff  GW_Temperatur  Phosphat   Sediment  \
Sauerstoff          8.209471      -0.862375 -0.029748   0.141498
GW_Temperatur      -0.862375       4.440565  0.029929  -0.013659
Phosphat           -0.029748       0.029929  0.011623  -0.006913
Sediment            0.141498      -0.013659 -0.006913   0.648204
Anzahl_Individuen  97.744211      16.880526 -2.753158  17.569737

                  Anzahl_Individuen
Sauerstoff                97.744211
GW_Temperatur             16.880526
Phosphat                  -2.753158
Sediment                  17.569737
Anzahl_Individuen      22199.736842
```



Aufgabenbesprechung

- Korrelation O2_Individuen = 0.22
- Alle Korrelationen:

```
corr_matrix = data.corr()
print(corr_matrix)
```

```
                  Sauerstoff  GW_Temperatur  Phosphat  Sediment  \
Sauerstoff          1.000000      -0.142830 -0.096302  0.061339
GW_Temperatur      -0.142830       1.000000  0.131738 -0.008051
Phosphat           -0.096302       0.131738  1.000000 -0.079638
Sediment            0.061339      -0.008051 -0.079638  1.000000
Anzahl_Individuen   0.228960       0.053764 -0.171394  0.146466

                  Anzahl_Individuen
Sauerstoff                 0.228960
GW_Temperatur              0.053764
Phosphat                  -0.171394
Sediment                   0.146466
Anzahl_Individuen          1.000000
```

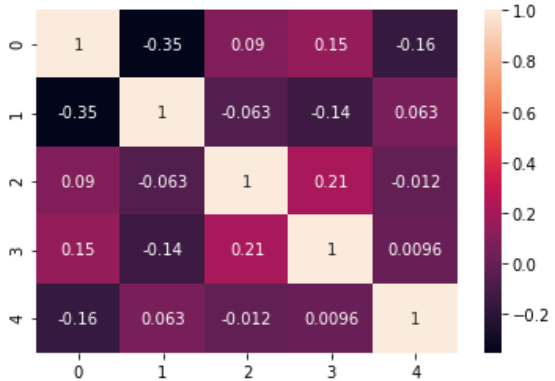
- Korrelation Pearson O2_GWT: r = - 0.14, p-Wert = 0.39
- Korrelation Spearman O2_GWT: r = - 0.06, p-Wert = 0.7

Aufgabenbesprechung

- Vergleich von drei Korrelationskoeffizienten

```
r_pear, p_pear = stats.pearsonr(data['Sediment'],data['Anzahl_Individuen'])
r_spear, p_spear = stats.spearmanr(data['Sediment'],data['Anzahl_Individuen'])
r_tau, p_tau = stats.kendalltau(data['Sediment'],data['Anzahl_Individuen'])
print (r_pear, p_pear, r_spear, r_tau, p_tau)
```

```
0.14646557616592315 0.3736111785965318 -0.07098288249384602 -0.05154355799732965 0.6701212825532348
```



Literatur

- Trauth (2015) MATLAB Recipes for Earth Sciences (4th Ed.), Springer
- Koch et al. (2020) Groundwater fauna in an urban area: natural or affected?, Hydrology and Earth System Sciences Discussions, https://hess.copernicus.org/preprints/hess-2020-151/

Nützliche Weblinks:

- Copluas: https://twiecki.io/blog/2018/05/03/copulas/


рахмат
danke
謝謝
teşekkür ederim
Баярлалаа
спасибо
dank je
blagodaram
thank you
gracias
tapadh leat
хвала
hvala
mauruuru
bedankt
dziękuję
mochchakkeram
go raibh maith agat
obrigado
sagolun
sukriya
kop khun krap
arigatō
takk
dakujem
ありがとう
grazie
terima kasih
мерси
merci
감사합니다