Für Mensch & Umwelt

Umwelt Bundesamt

Analysenergebnisse aus dem Internet?

Automatisierte Sammlung und Auswertung öffentlich verfügbarer Informationen zur Trinkwasserqualität in Deutschland

Ergebnisse des BMG-Kurzforschungsprojektes TriSto

Leon Saal
Aki Sebastian Ruhl
Fachgebiet II 3.3 / Wasseraufbereitung

50 Jahre Umweltbundesamt 1974–2024

TriSto

Gefördert durch:

Bundesministerium für Gesundheit

aufgrund eines Beschlusses des Deutschen Bundestages

- Kurzprojekt finanziert durch das BMG
- Laufzeit Dezember 2021 – April 2022 (5 Monate)

„Evaluierung der Beschaffenheit von Trinkwasser auf Grundlage öffentlich verfügbarer Informationen zu Stoffkonzentration in großen Wasserversorgungsgebieten Deutschlands“

HINTERGRUND:

- Aktuell kein Überblick über Konzentrationsniveaus der geregelten (und weiterer) Wasserinhaltsstoffe verfügbar
- Wichtig für:
  - Information der Öffentlichkeit
  - Anpassung von Grenzwerten

ZIELE:

- Sammlung der Konzentrationswerte
- Bewertung der Verfügbarkeit und Nutzerfreundlichkeit

Vorgehen

GRUNDLAGE:

- Gesuchte Informationen liegen dezentral vor
- Viele Wasserversorgungsunternehmen (WVU) unterhalten Websites mit Informationen

SCHEMATISCHES VORGEHEN:

Vorgehen: Datensammlung “web scraping”

1. Emulieren von erwartetem Nutzer*innenverhalten

- Suchbegriff: Gemeindenamen aus größten Wasserversorgungsgebieten (WVG) in DE, z.B. Berlin
- WVG wurden absteigend nach versorgten Bürger*innen durchgegangen

2. Abruf der Links zu ersten vier Suchergebnissen
3. Download von PDF-, Excel- und Bilddateien, sammeln von Metadaten

Vorgehen: Aufbereitung der Daten

- Lokal gespeicherte Dateien wurden:

1.Auf Trinkwasserqualitätsdaten hin geprüft und Metadaten aufgenommen

1.Tabellarische Daten extrahiert

- Anhand von Dateiname
- Dokumenttext auf Parameternamen hin durchsucht

- Wenn Dokument nicht eingescannt
- Seitenanzahl < 25

1.Daten in Datenbank eingespeist

1.Daten mit Parameter und Einheitsspalte (Index)

- Duplikate in den Tabellen
- Festlegung von Parameter und Einheitsspalte
- Ausgeschlossen

- Weitere Datenaufreinigung

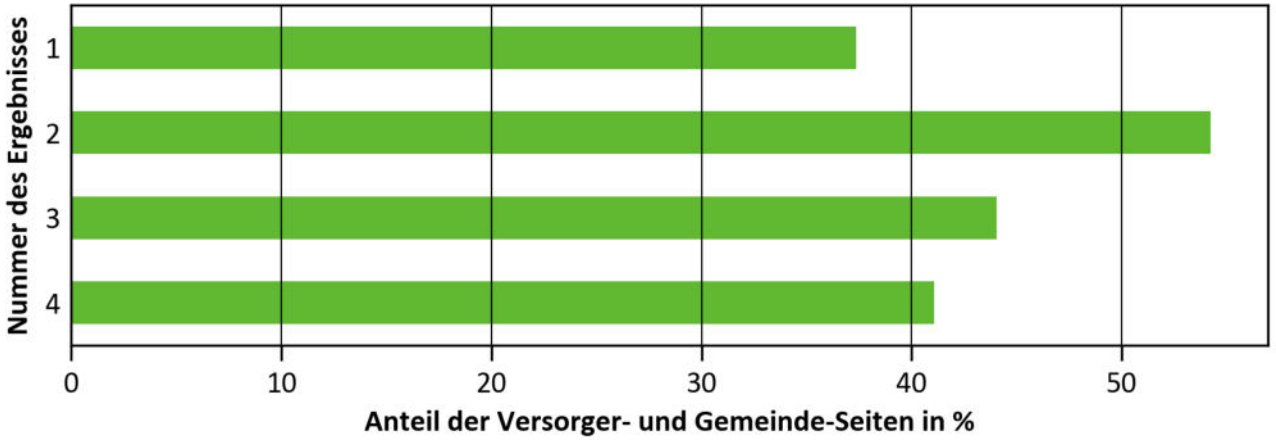
Spalten ohne Messwerte ausgeschlossen

In einheitliche Einheiten umgerechnet

- Laufende Nummerierung
- Spalten mit Grenzwerten

Ergebnisse: Datensammlung

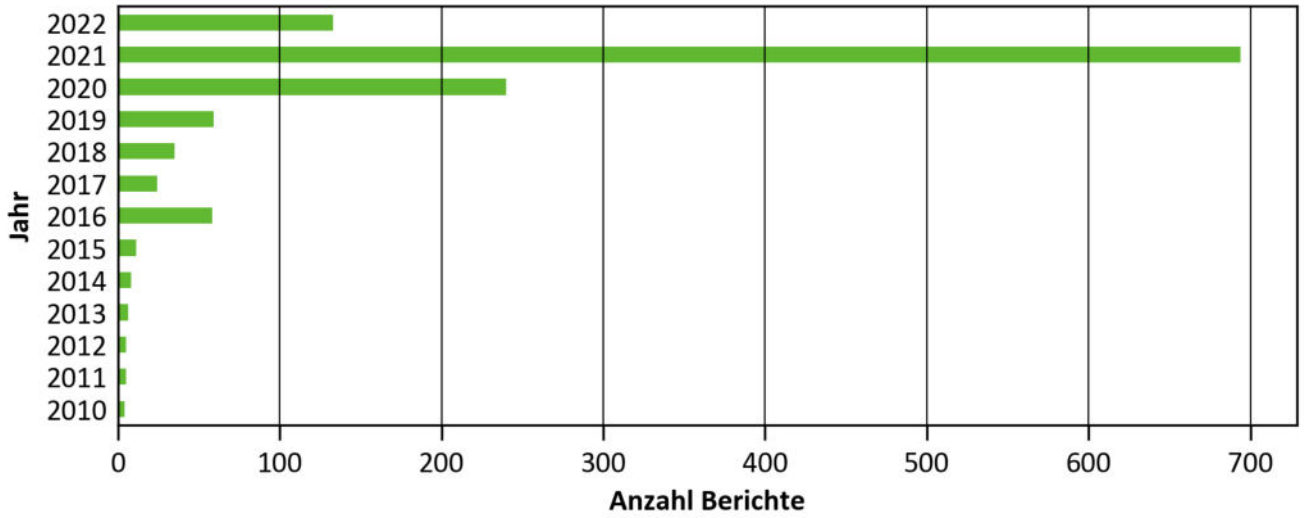
- Suche für 701 WVG, die 53,9 Mio. EW mit täglich 8,41 Mio. m³ versorgen
- Für 1168 von 2096 Gemeinden wurden Messwerte gefunden (55 %, in 510 WVG)
- Prominenz bei Suche:



- Erste Einträge meist durch kommerzielle Anbieter von Wassertests dominiert (4.033 von 9.272 Suchergebnissen 43 %)

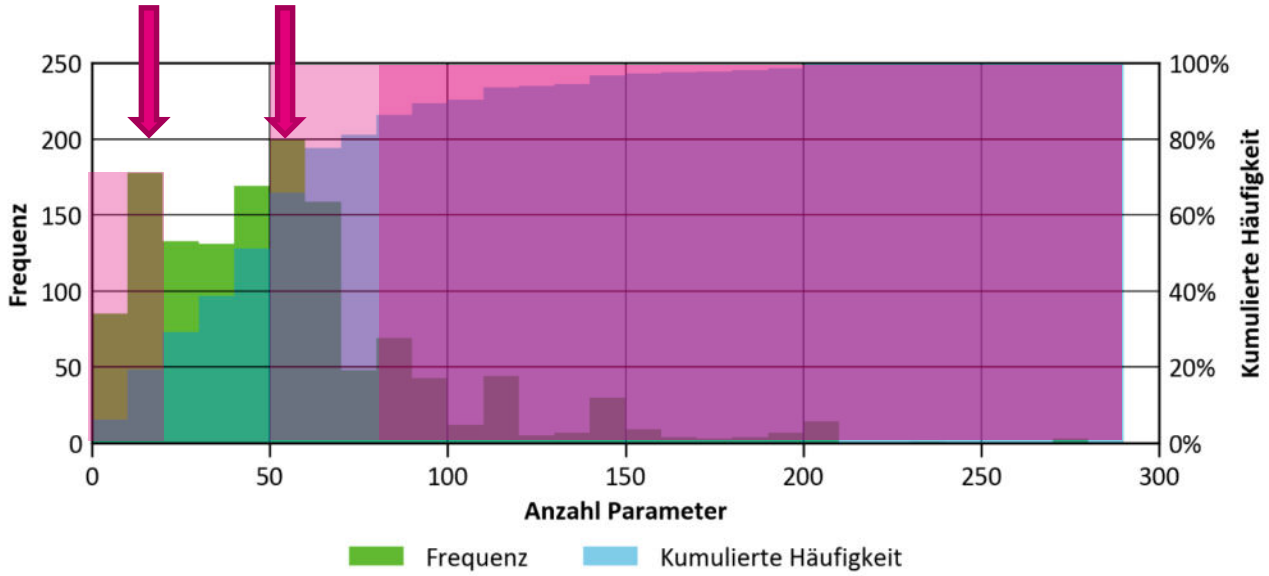
Ergebnisse: Aktualität der Daten

- Datensammlung erfolgte an drei Tagen Anfang 2022 (17.02, 18.02, 7.03)
- Großteil der Berichte vom Jahr vor Sammlung, viele von 2020 und einige neue aus Jahr der Sammlung



Ergebnisse: Parameter pro Bericht

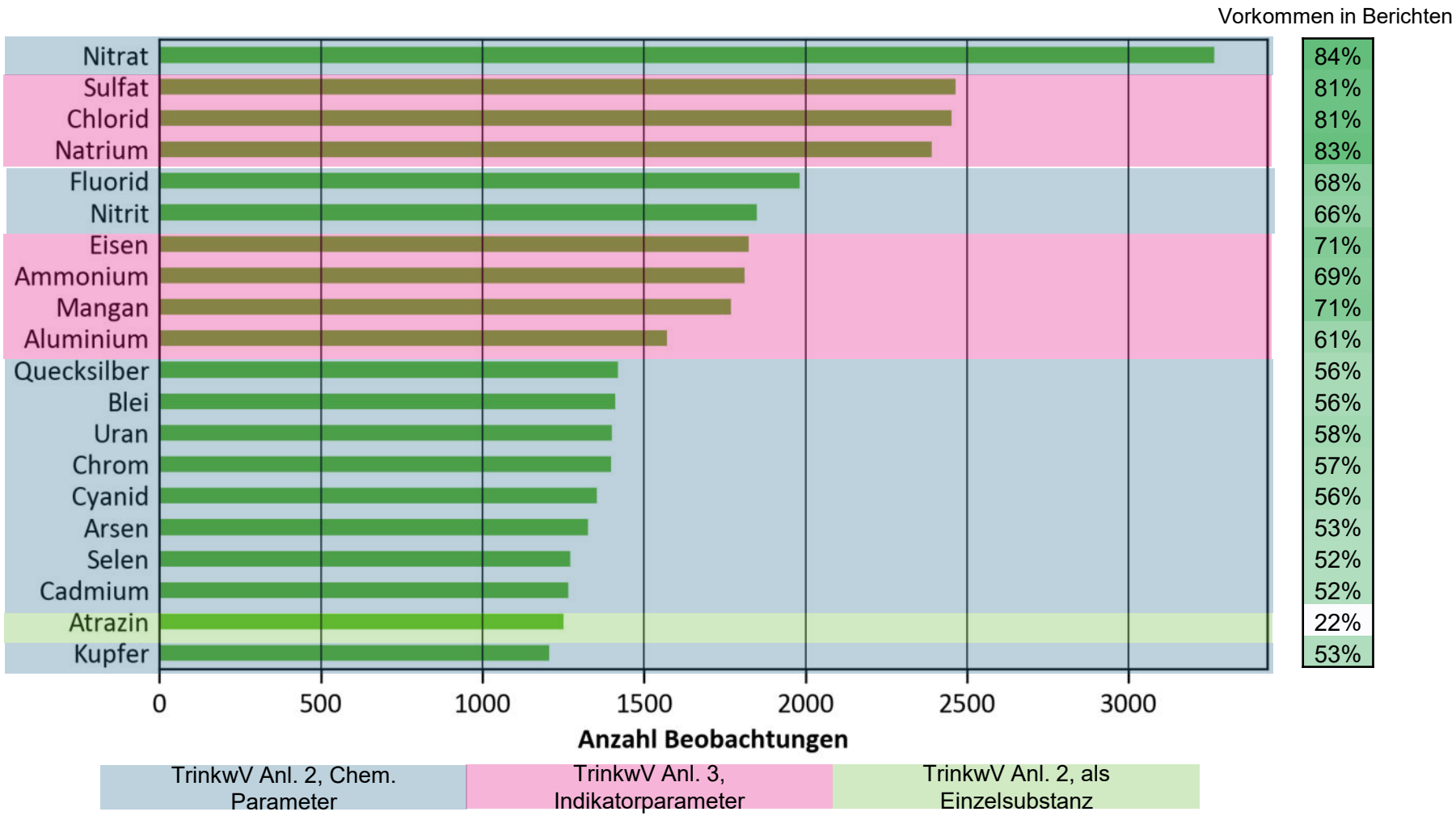
- Variierende Anzahl berichteter Parameter
- Häufig: 10-20, 50-60
- 20 % der Berichte enthalten weniger als 20 Parameter
- 50 % berichten mehr als 50
- 20 % der Berichte führen mehr als 80 Parameter



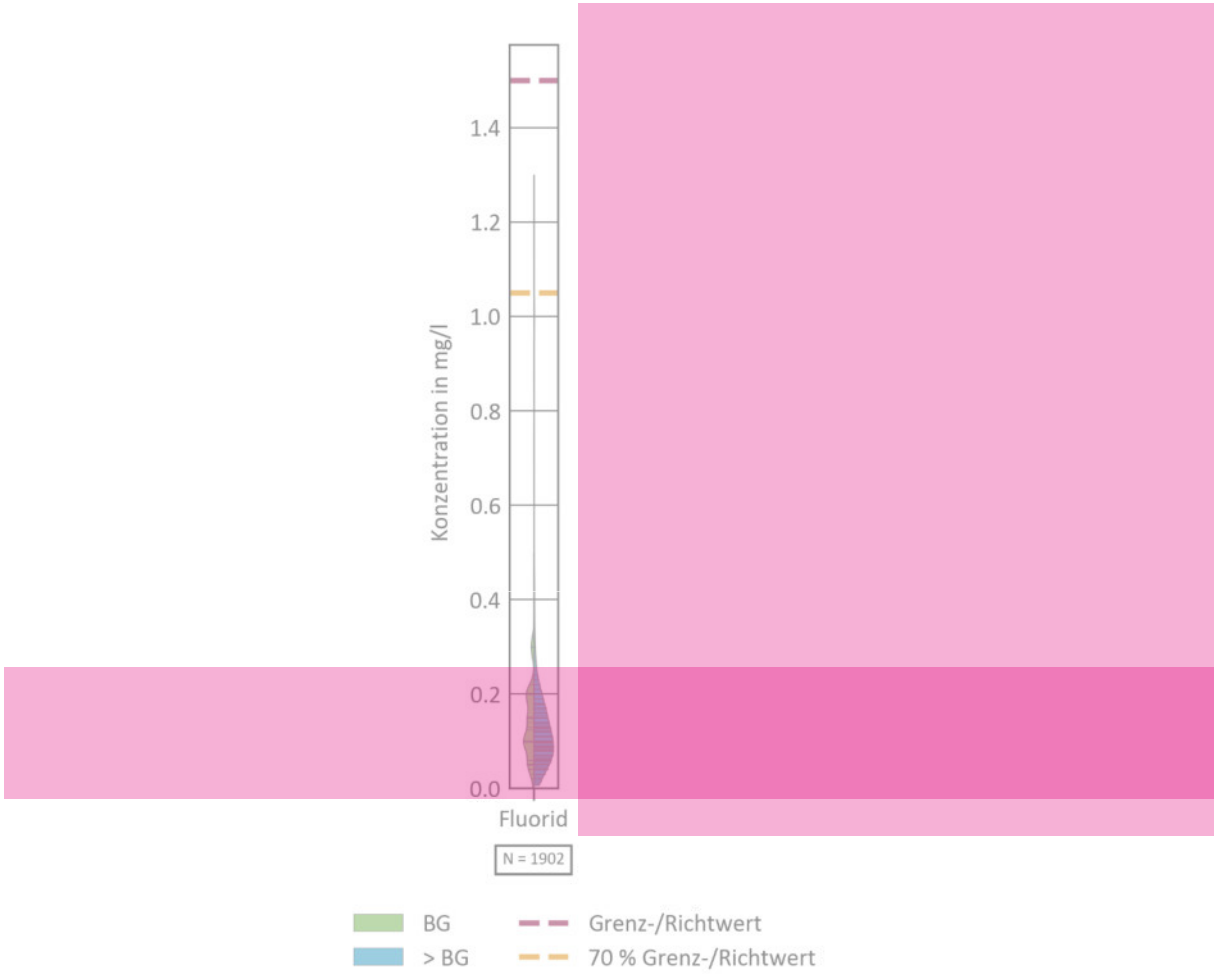
Anzahl Parameter nach TrinkwV

- Mikrobiologische Parameter: 2
- Chemische Parameter: 27
- Indikatorparameter: 20

Ergebnisse: 20 häufigste Parameter mit Grenzwert

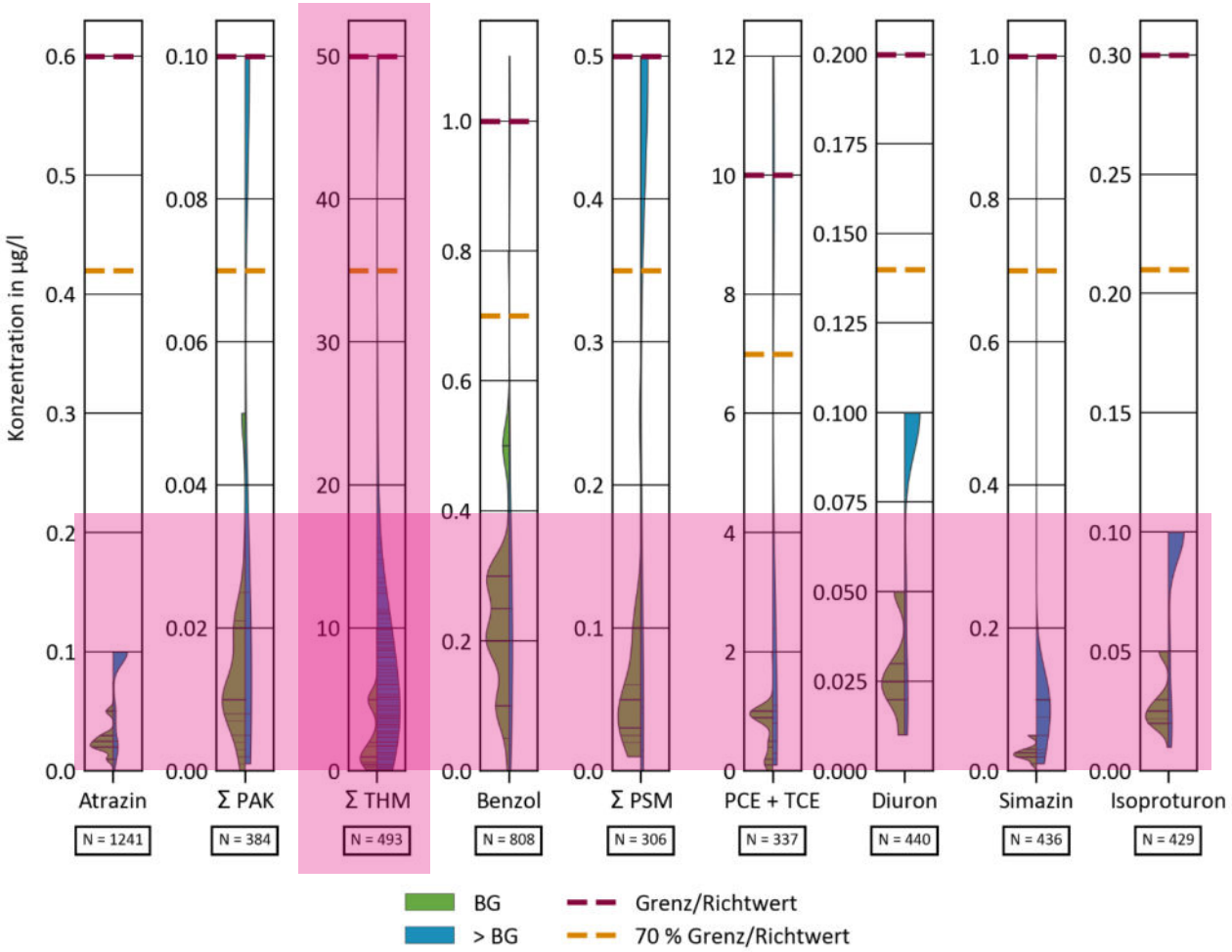


Ergebnisse: Werteverteilung der häufigsten Parameter



- Großteil Werte < 20% des Grenzwertes → Ausnahme Nitrat
- Parameter mit Grenzwert < 1 mg/l häufiger < BG als darüber

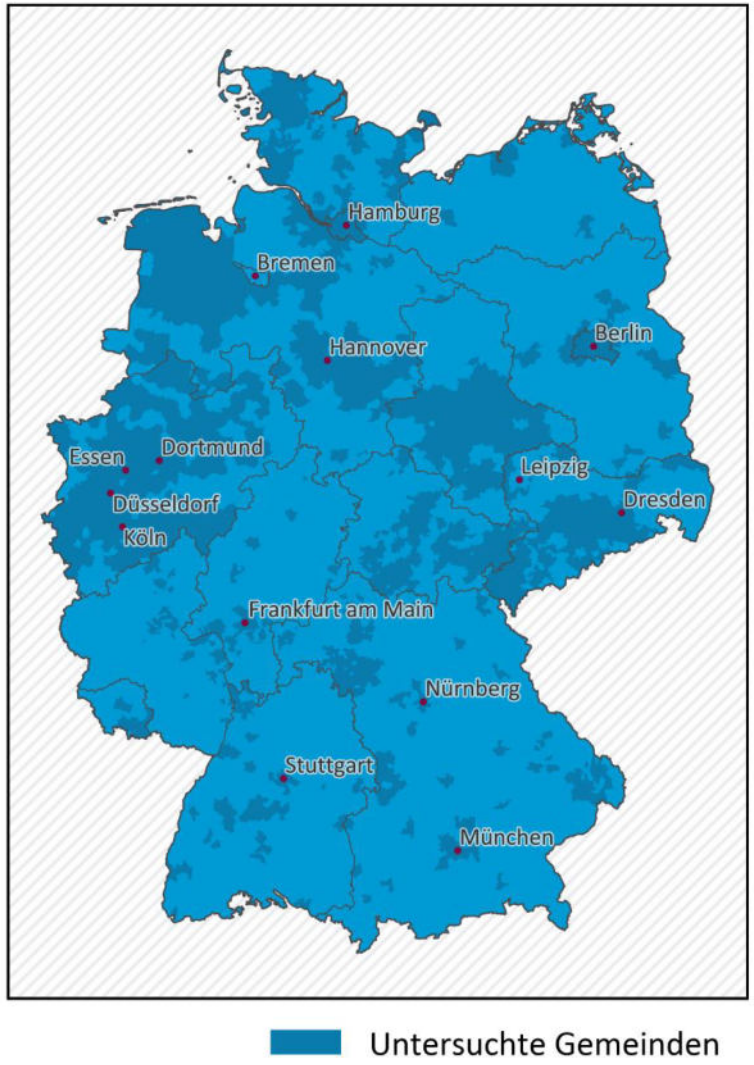
Ergebnisse: Werteverteilung der häufigsten Parameter



- Organische Spurenstoffe viel häufiger < BG als darüber
- Ausnahme THM (Desinfektions- und Oxidationsnebenprodukte)
- BG << 70 % Grenz-/Richtwert

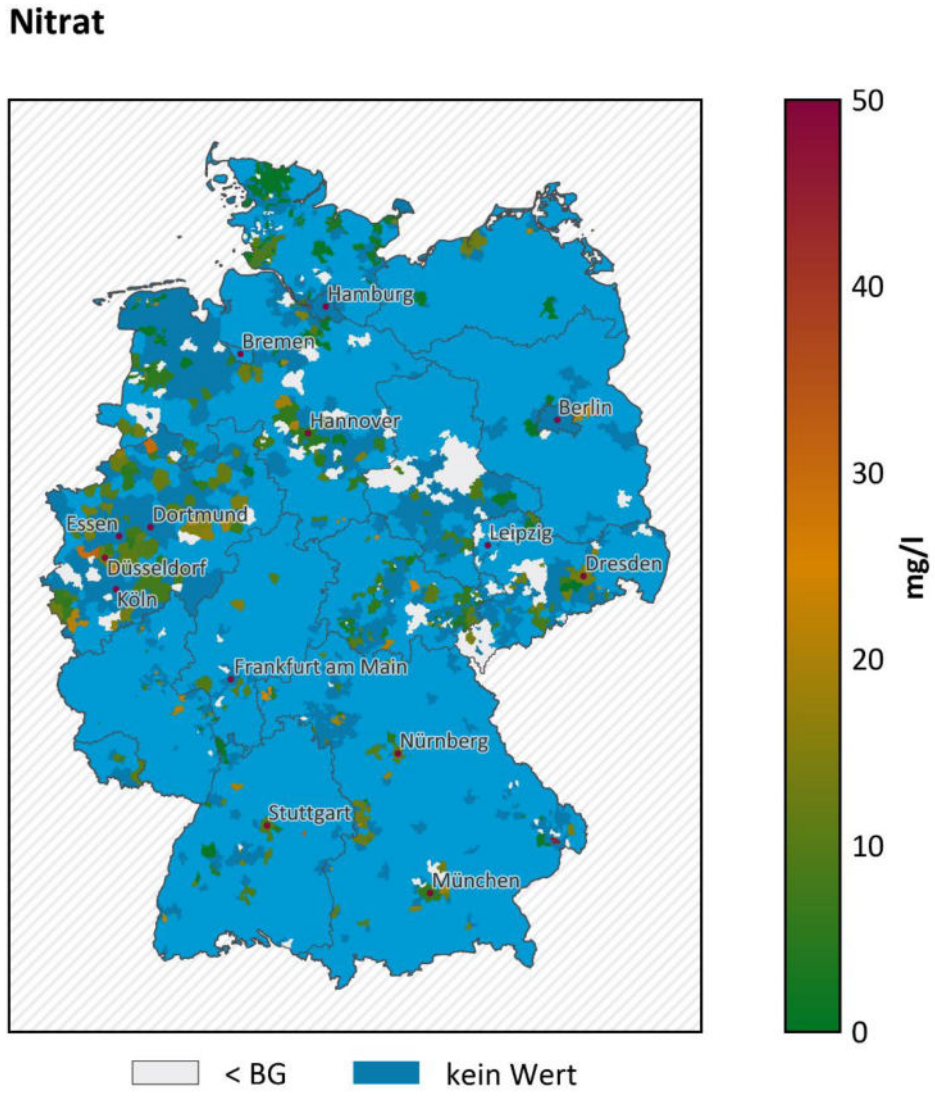
Ergebnisse: Geografische Verteilung häufiger Parameter

- Untersuchte Gemeinden haben Fläche von 95.200 km² (26,6 % von Gesamtfläche von 357.580 km²)
- > 50 % der Gemeinden (nach Fläche und EW) in NRW, NDS und BY
- Flächenabdeckung ungleichmäßig



Ergebnisse: Geografische Verteilung häufiger Parameter

- Durch Auswahl der zu untersuchten WVG
  Bevorzugung städtischer Gebiete
- Aber: Erhöhte Ausbringung von Gülle und erhöhte Konzentration eher ländlich



Zusammenfassung

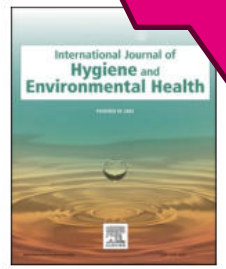
Open Access

International Journal of Hygiene and Environmental Health 255 (2024) 114295

Contents lists available at ScienceDirect

International Journal of Hygiene and Environmental Health

journal homepage: www.elsevier.com/locate/ijheh



Check for updates

Automated scraping and analyses of drinking water quality data

Leon Saal [a,b], Aki Sebastian Ruhl [a,b,*]

[a] German Environment Agency, Section II 3.3, Schichauweg 58, 12307, Berlin, Germany
[b] Technische Universität Berlin, Chair of Water Treatment, Sekr. KF4, Straße des 17. Juni 135, 10623, Berlin, Germany

(Fokus auf Städte)

1.

Saal, L. & Ruhl, A. S. Automated scraping and analyses of drinking water quality data. *International Journal of Hygiene and Environmental Health* **255**, 114295 (2024). https://doi.org/qbq9

Ausblick

- Problemstellung geeignet für KI-Anwendung. Potentiell:
  - Bessere Tabellenauswertung
  - Bessere Erkennung von Parametern

- Wiederholte Datensammlung in regelmäßigen Abständen (z.B. jährlich)
  - Beobachtung von Digitalisierung
  - Beobachtung von Umsetzung neuer Überwachungsregularien

SHAPTH
Schnittstellenharmonisierung und Austauschplattform Trinkwasserhygiene

[…] soll eine Harmonisierung der in Deutschland genutzten Datenaustauschformate für (Trink-)Wasseranalysen und damit verbundener Daten und Kataloge erfolgen. Ein allgemein verwendbares Datenformat soll definiert und darüber hinaus eine einheitlich verfügbare Datenaustauschplattform implementiert werden. (https://shapth.info/)

Vielen Dank für Ihre Aufmerksamkeit

leon.saal@uba.de
akisebastian.ruhl@uba.de