

EGU25-21589, updated on 29 Jan 2026

<https://doi.org/10.5194/egusphere-egu25-21589>

EGU General Assembly 2025

© Author(s) 2026. This work is distributed under the Creative Commons Attribution 4.0 License.



## Applicability of imputation methods for enhancing density of long term hydrogeochemical data — Differences and constraints of conventional and machine learning-based approaches

**Danilo Veskov**<sup>1</sup>, Dirk Antunovic<sup>2</sup>, Björn Droste<sup>2</sup>, and Dr. Ferry Schipperski<sup>1</sup>

<sup>1</sup>Fachgebiet für angewandte Geochemie, Institut für ang. Geowissenschaften, Technische Universität Berlin, 10587 Berlin, Deutschland

<sup>2</sup>Qualitätsüberwachung Wasser, Wasserwerke/Wasserwirtschaft, Stadtwerke Düsseldorf AG, 40589 Düsseldorf, Deutschland

Despite adequate water availability, Germany faces a widespread need to optimize the sustainable use of groundwater due to high water utilization rates. Furthermore, being a main source for drinking water, groundwater needs profound and future-proof protection. To address these challenges, aquifer management practices must be improved for greater efficiency in order to maintain the long-term availability of good drinking water quality.

Statistical analysis of hydrogeochemical data offers valuable insights into the functioning of groundwater systems, the identification of dominant processes within aquifers, and the detection of contaminant input sources. Although long-term data is often available, the variable structure of these datasets frequently poses challenges for immediate statistical analysis. Data sparsity caused by the integration of datasets with differing parametric and temporal resolutions (e.g., data from scientific research programs versus routine monitoring programs by governmental water suppliers) poses a problem for statistical evaluation methods sensitive to data density (e.g., Principal Component Analysis). Instead of the rigorous deletion of time steps and/or parameters in cases, where data density is critical for the selected evaluation method, preprocessing by imputation can reduce the loss of valuable information.

This study demonstrates the applicability, limitations and distinctions of common script-based imputation methods for enhancing the density of long-term hydrogeochemical data. Two datasets of groundwater from two different drinking water protection areas in Germany (Düsseldorf and Dormagen, 2000–2023) and a third dataset from the Rhine River (dividing both protection areas, 1990–2023) were evaluated (provided by the Stadtwerke Düsseldorf AG within the framework of the research project iMolch, a collaborative project of the funding measure LURCH). The evaluations span conventional imputation methods to modern machine-learning approaches, while indicating an emerging new potential for the re-assessment of historical data through the utilization of recently available machine-learning algorithms. Nonetheless, data imputation must be applied cautiously, as it carries the risk of introducing non-representative data values, particularly when conducted without thorough understanding of the data structure, internal

dependencies, and the imputation mechanism. Additionally, both the effectiveness of the imputation and the preservation of the data's representativeness should be strictly verified post-application. Therefore, the results highlight methods-specific constraint differences, offering practical, Python-based recommendations for efficient hydrogeochemical data imputation.

By enhancing data density while preserving representativeness, this work contributes to addressing the broader challenge of optimizing the sustainable groundwater use and safeguarding water resources under increasing anthropogenic pressures.