UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

DEPARTMENT OF EARTH AND ENVIRONMENTAL SCIENCES

EGU General Assembly 2020

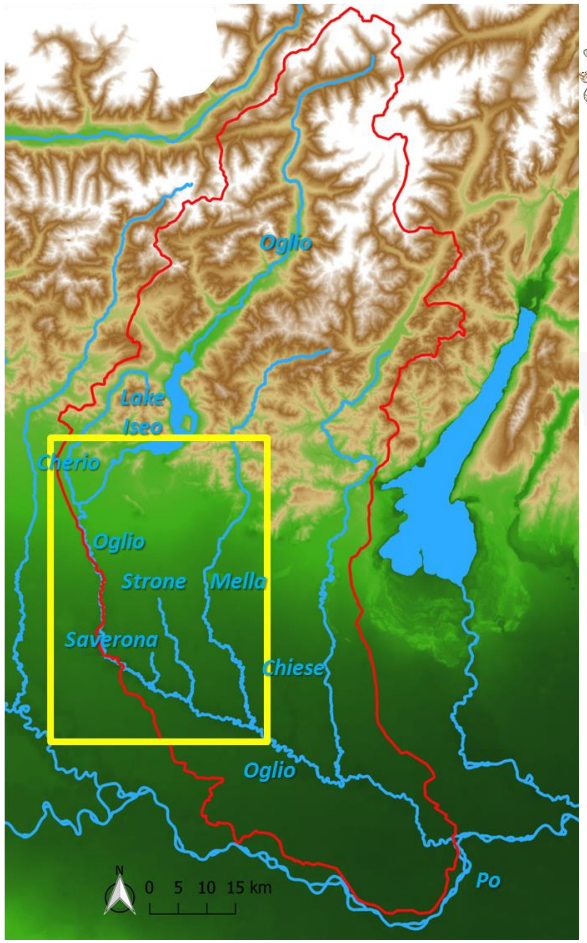
Positive Matrix Factorization and GIS approach to perform data mining on groundwater and surface water quality dataset.

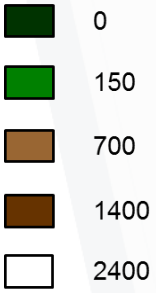
Zanotti, C.[1], Rotiroti, M. [1], Fumagalli, L. [1], Stefania, G.A. [1], Canonaco, F. [2], Stefenelli, G. [2], Prévôt, A.S.H. [2], Leoni, B [1], Bonomi, T. [1].

[1] Department of Earth and Environmental Sciences, University of Milano-Bicocca, Piazza della Scienza, 1, 20126 Milano, Italy. Email: c.zanotti@campus.unimib.it

[2] Laboratory of Atmospheric Chemistry, Paul Scherrer Institute, 5232 Villigen-PSI, Switzerland

INTRODUCTION





fondazione cariplo
grant n. 2014-1282

Main aims:

- To identify the main hydrochemical features of either groundwater and surface water (lake, river and springs) and, possibly, the processes that influence (or govern) them.
- To understand the relations (if any) between the chemistry of groundwater and surface water.

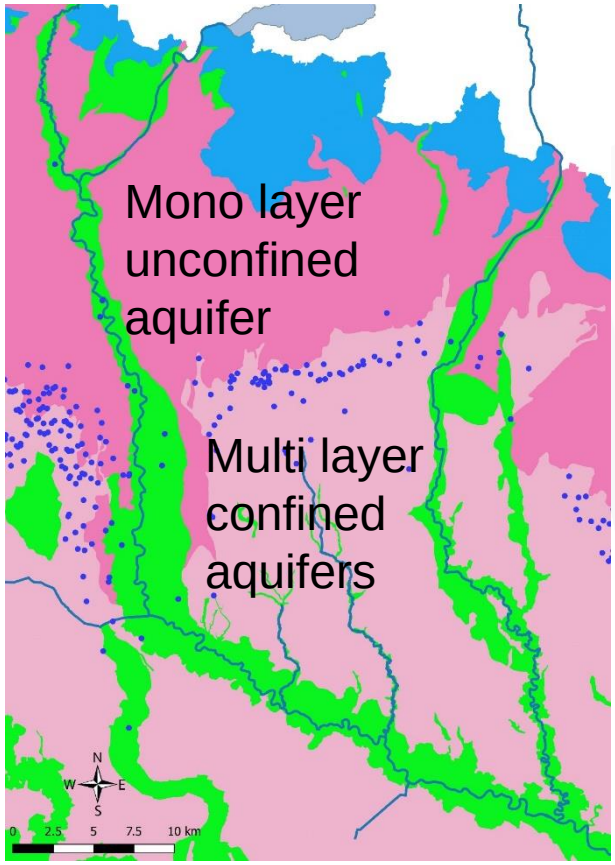
Methodology:

- Application of Positive Matrix Factorization (PMF)

Oglio river catchment area: 6360 km^2
Study area: 1960 km^2

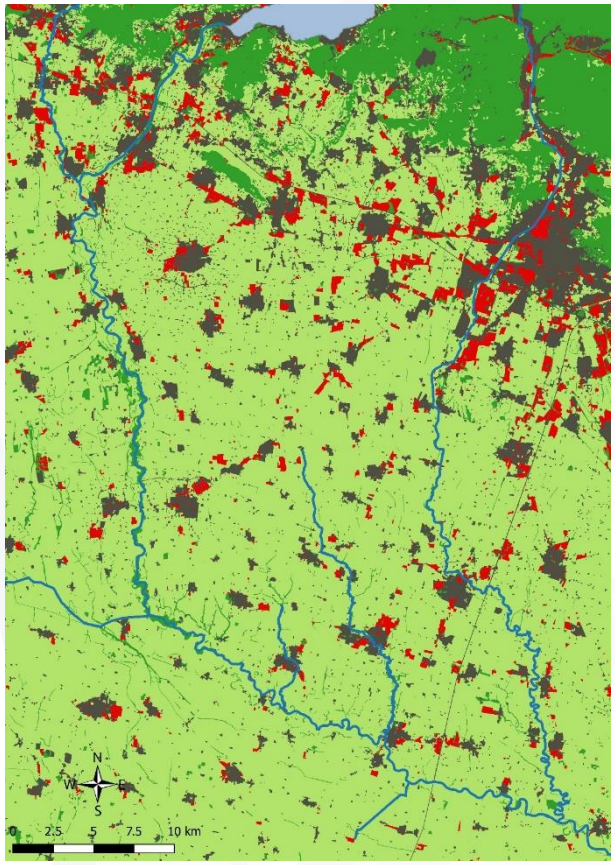
1 – STUDY AREA

Geology



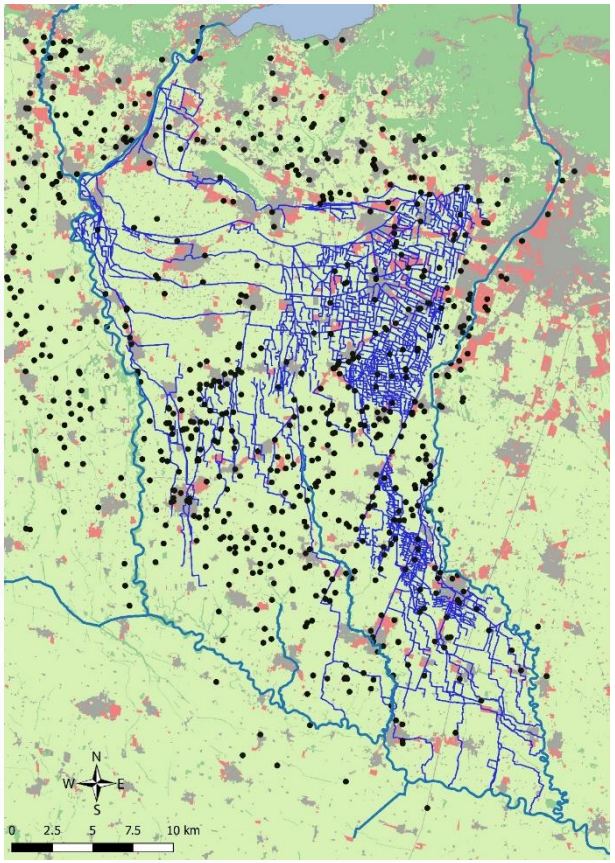
- Moraine
- Higher Plain
- Low Plain
- River Valley
- springs

Landuse



- Agriculture
- Urban
- Industry
- Natural

Irrigation sources



- Irrigation channels
- Irrigation wells

DATASET

4 field surveys:

- February 2016
- June 2016
- September 2016
- March 2017.

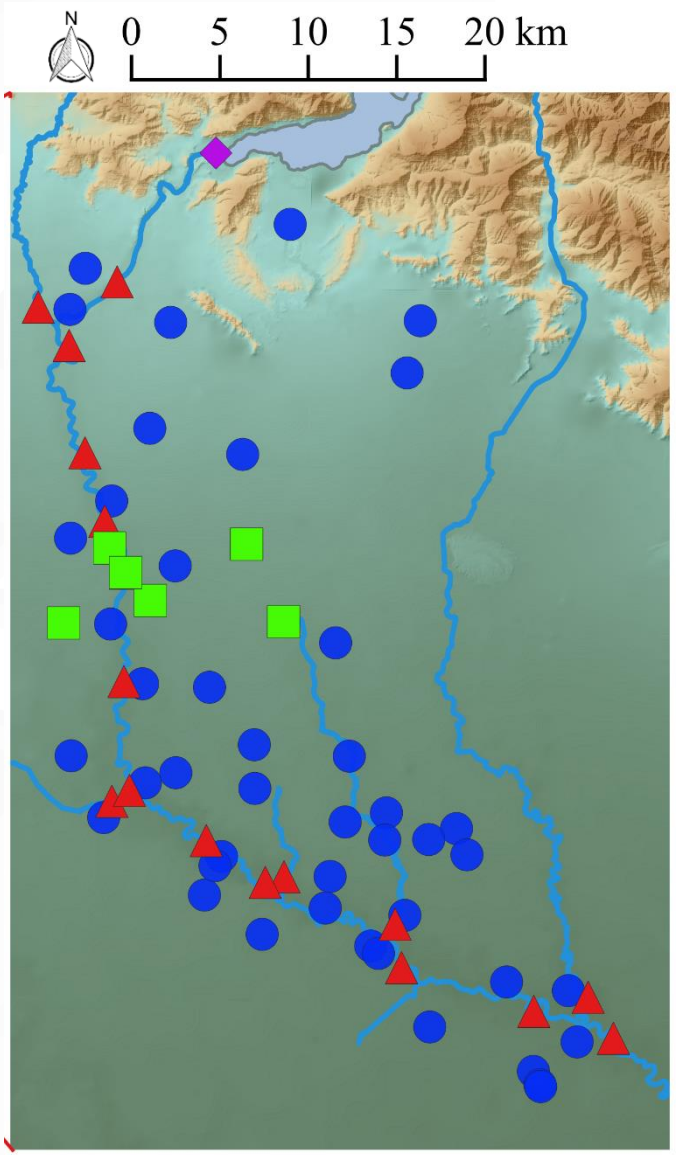
Measured in water samples from:

- 44 wells
- 1 lake station
- 17 river stations
- 6 springs

Chemical variables (14):

- Dissolved O_2
- Ptot
- Major ions (HCO_3, Ca, Mg, Na, K, Cl, SO_4, NO_3, NH_4)
- Trace elements (As, Fe, Mn)

For the validation: Water Isotopes



MATERIALS AND METHODS

Problem related to the reduction the original data-set into one of lower dimensions to detect "hidden" information and explain the variability of the measured variables.
Usually this kind of problem is solved with Factor Analysis, but it presents some limitations:

Factor Analysis (FA)

- Well known and standardized
- Widely used in hydrology

- Factors grouping variables with both positive and negative correlation
- It is sometimes difficult to interpret and disseminate
- No way to treat data uncertainty

Positive Matrix Factorization (PMF) is a multivariate analysis aimed at source identification and apportionment, specifically designed to cope with environmental data and manage their uncertainty and distributions. Here it is tested the effectiveness of PMF as a tool to perform data mining and define hydrochemical features of groundwater and surface water and to understand their relationship.

Positive Matrix Factorization (PMF)

- Positivity constraint
- Data uncertainty

How does PMF work in hydrology?

MATERIALS AND METHODS

Positive Matrix Factorization (PMF)

PMF (Paatero and Tapper 1994) is a multivariate analysis in which the fundamental problem is to resolve the identities and contributions of different sources in a mixture.

$$x_{ij} = \sum_{k=1}^{p} g_{ik} f_{kj} + e_{ij} \qquad i = 1 \dots m; j = 1 \dots n; k = 1 \dots p$$

- NO orthogonality constraint
- 2 Positivity constraints:
  - the composition of predicted source must be positive (a source cannot have negative percentage of an element);
  - the predicted source contribution to each sample must all be positive (a source cannot contribute to a sample with a negative mass).
- The solution of the problem is found trough a weighted least squares approach, minimizing:

$$Q = \sum_{i=1}^{n} \sum_{i=1}^{m} \left(\frac{x_{ij} - \sum_{k=1}^{p} g_{ik} f_{kj}}{\sigma_{ij}} \right)^2$$

Data uncertainty

Allows to work with missing data and data below detection limit

Particularly suitable for environmental data

MATERIALS AND METHODS

Informatic tool:

Atmos. Meas. Tech., 6, 3649–3661, 2013
www.atmos-meas-tech.net/6/3649/2013/
doi:10.5194/amt-6-3649-2013

Atmospheric Measurement Techniques Open Access

SoFi, an IGOR-based interface for the efficient use of the generalized multilinear engine (ME-2) for the source apportionment: ME-2 application to aerosol mass spectrometer data

F. Canonaco, M. Crippa, J. G. Slowik, U. Baltensperger, and A. S. H. Prévôt
Paul Scherrer Institute, Laboratory of Atmospheric Chemistry, 5232 Villigen PSI, Switzerland

Correspondence to: A. S. H. Prévôt (andre.prevot@psi.ch)

1 – UNCERTAINTY ANALYSIS

Two major sources of uncertainty in PMF analysis:

- Rotational ambiguity (more than one solution can give the same Q)
- Random errors in the data

To asses PMF uncertainty:

- 100 PMF runs over bootstrapped data
- At each run factor profiles rotates within predefined range

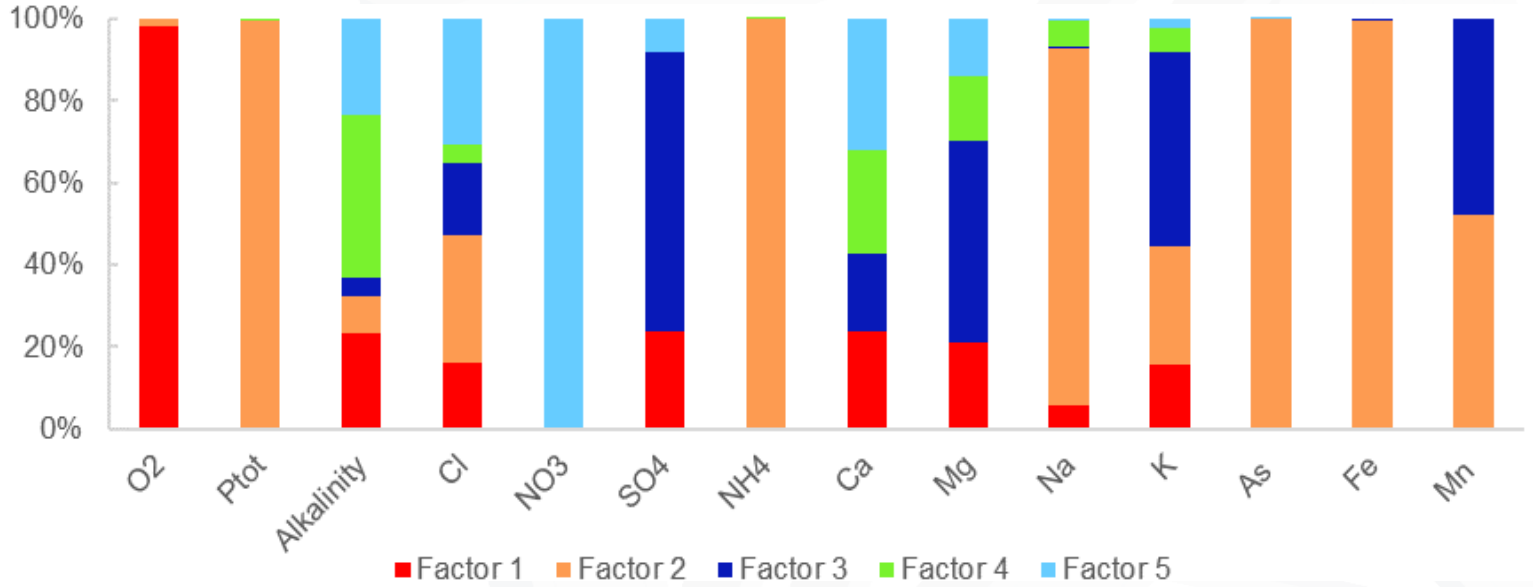
Results are expressed as average over the different runs and the uncertainty as the standard deviation.

At each run only 70% of the data are randomly resampled from the original dataset and used used to solve the model
This reduces the leverage of single data, especially extremes

1 – RESULTS – PMF factors profiles

PMF generated 5 Factors

Each one of them was associated to a specific process or feature of the system, based on thei factor profile or their spatial distribution



1- **Lake Iseo and Oglio river**: highly oxygenated water, with SO_4 and a small content of major ions such as HCO_3, Cl, Ca, Mg, Na, and K

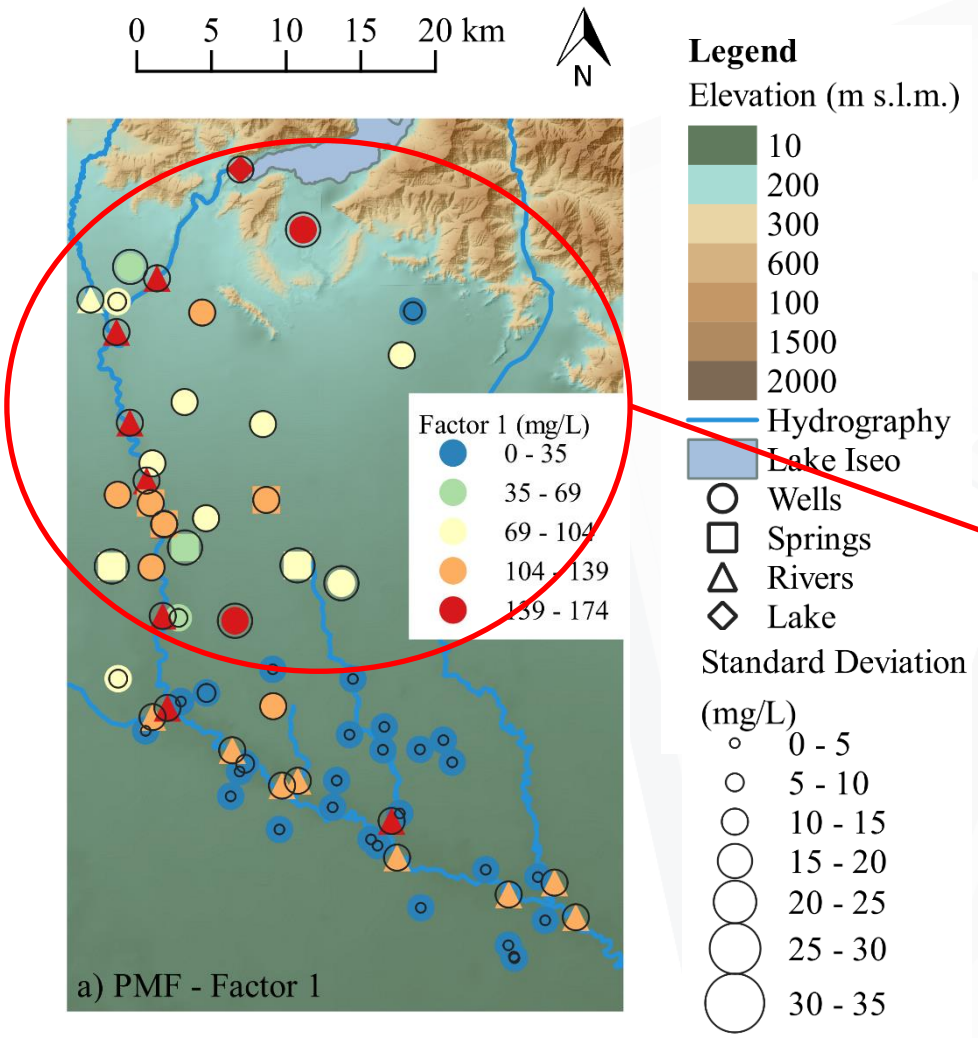
2- **Reducing condition - advanced stages**: As, Fe, P-tot and NH_4, furthermore it has a significant contribution of Mn

3- **Reducing condition - early stages**: mainly characterized by Mn and SO_4, with a contribution of the major ions

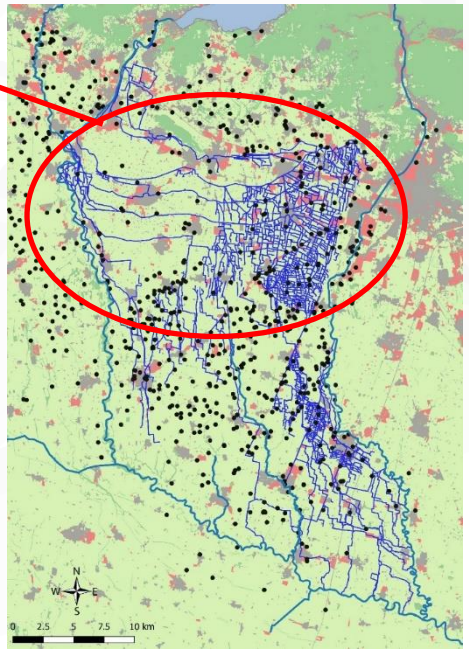
4- **Residence time**: only major ions such as HCO_3, Cl, Ca, Mg, Na, and K

5- **Anthropogenic impact:** NO_3 with contributions also in terms of Cl, SO_4, Ca and Mg.

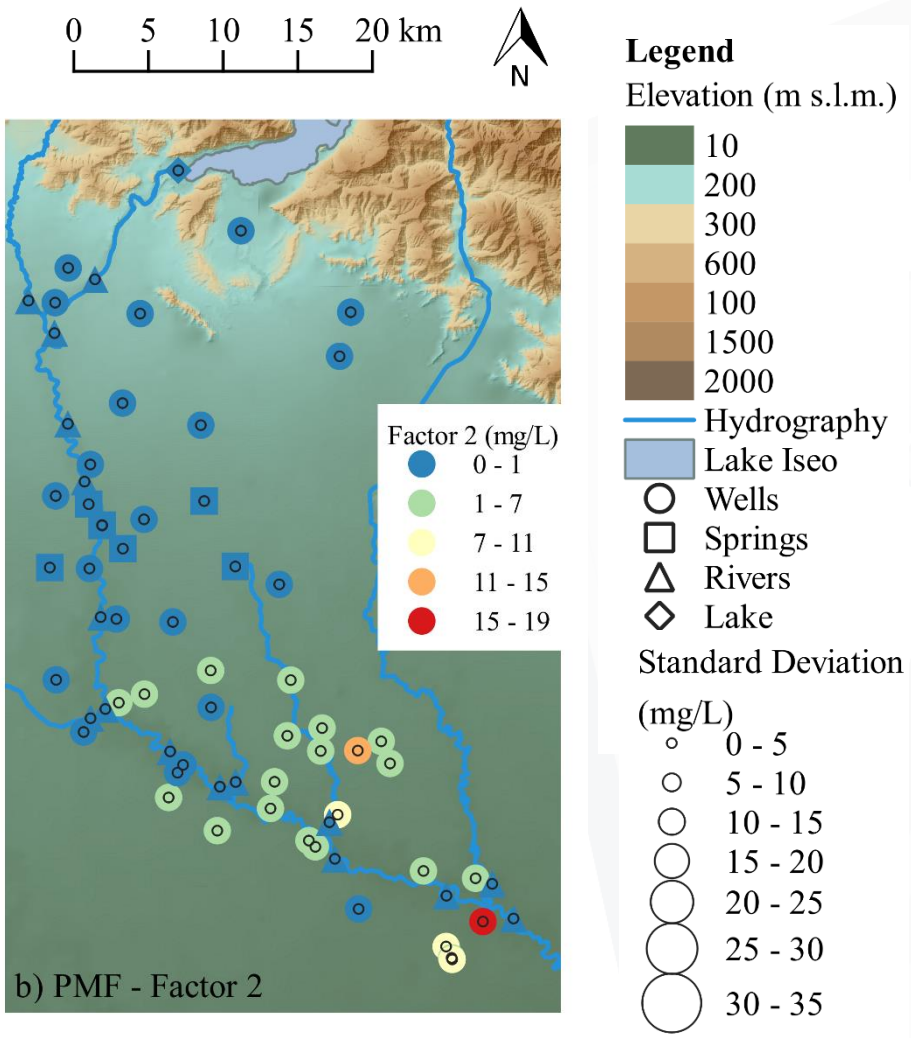
PMF Fac 1 – Lake Iseo and Oglio River



- Water from Lake Iseo flows into the Oglio river
- Higher values in the higher plain wells due to irrigation processes: channels collect water from the Oglio river and spread it over the fields from where it can reach the aquifer

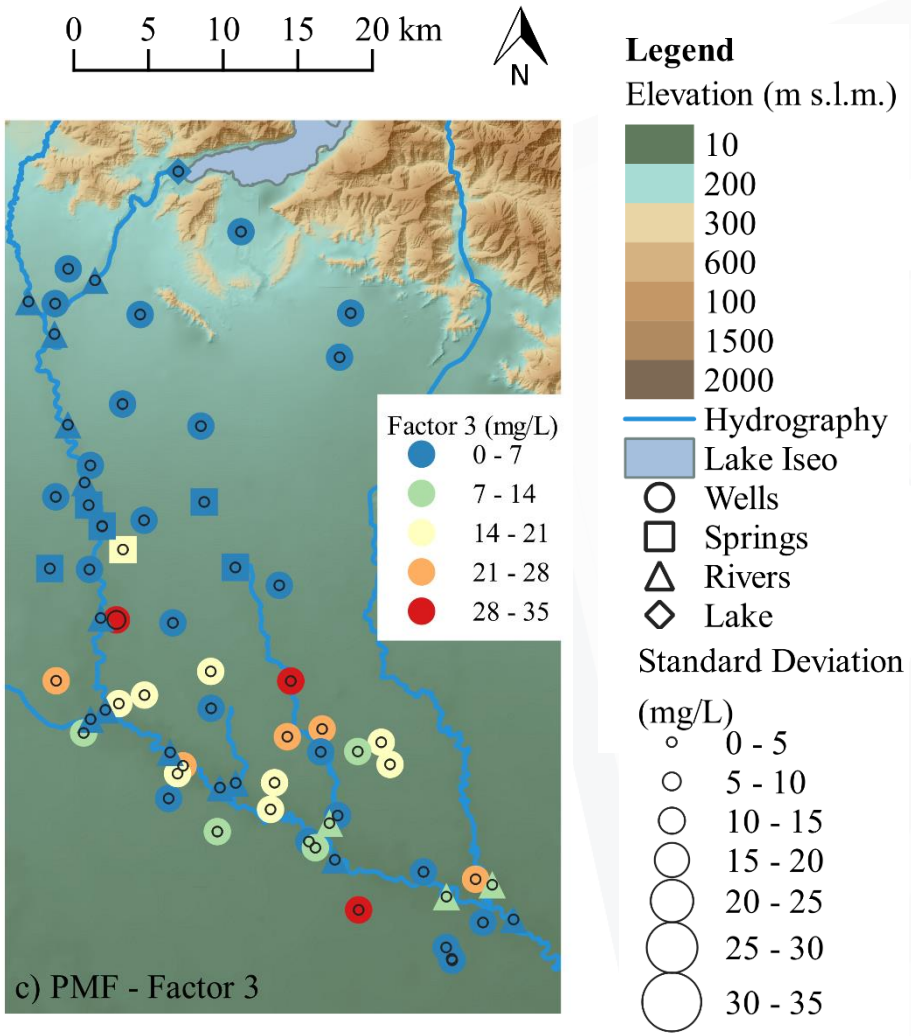


PMF Fac 2 – Advanced Red. conditions



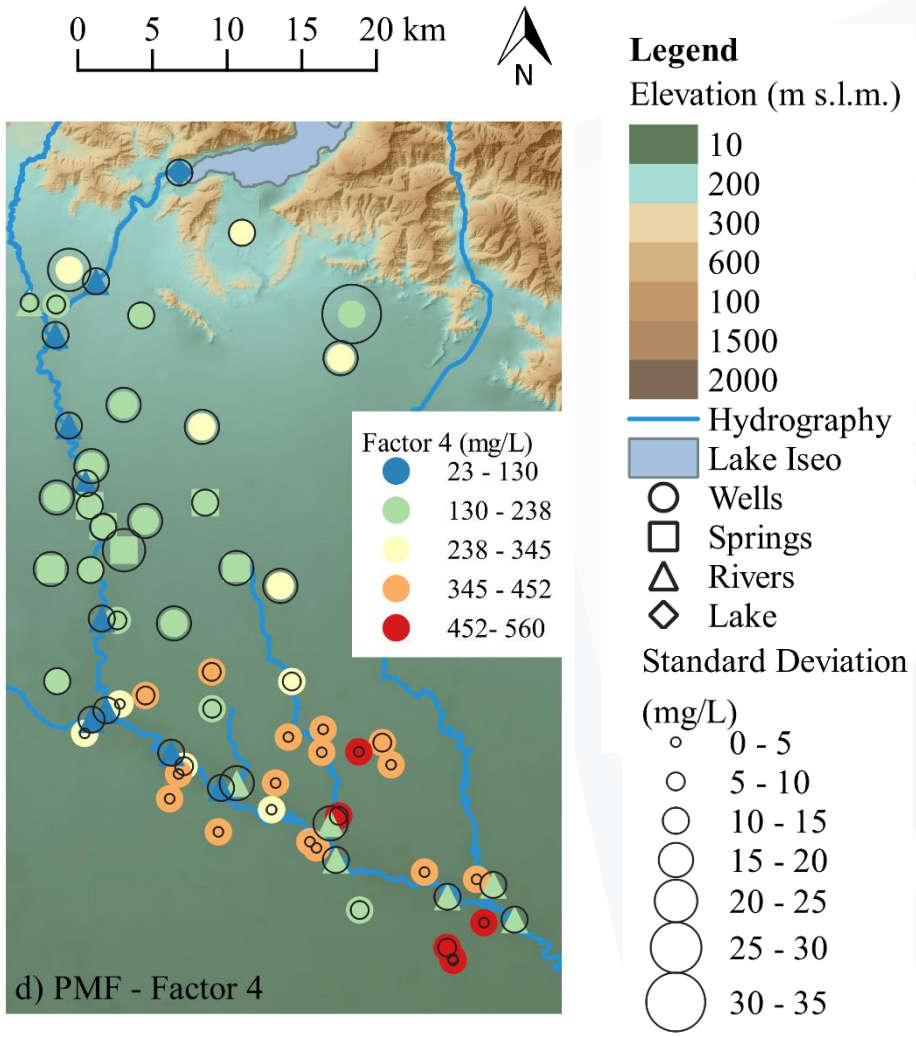
- Sampling points with higher values are wells in the lower plain, where As pollution is stronger
- Groundwater samples in the higher plain, which is unconfined, and surface water samples have no contribution of this factor.

PMF Fac 3 – Early Red. conditions



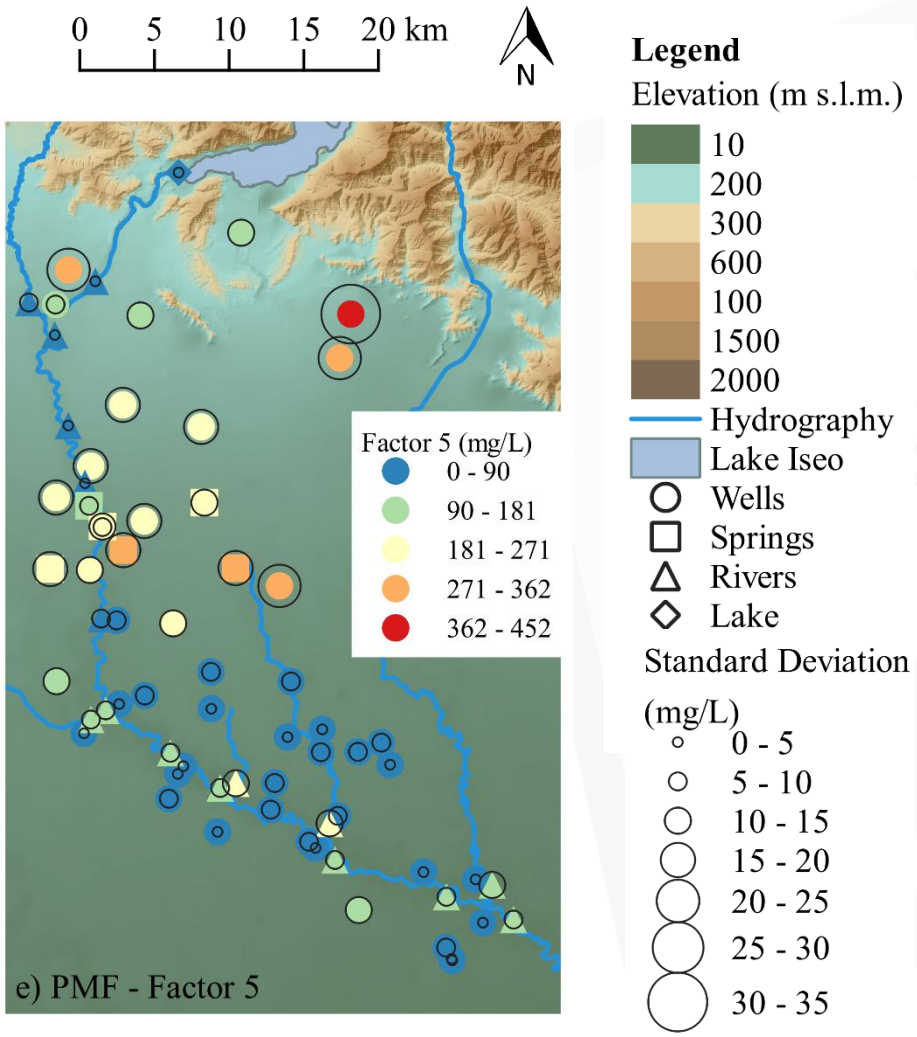
- sampling points with the highest values at the transition between higher and lower plain aquifers
- here groundwater is passing from the oxidising conditions of the higher plain to the reducing conditions of the lower plain, so Mn-oxide reduction can be favoured.
- groundwater samples in the higher plain and surface water samples have no contribution of this factor.

PMF Fac 4 – Residence time



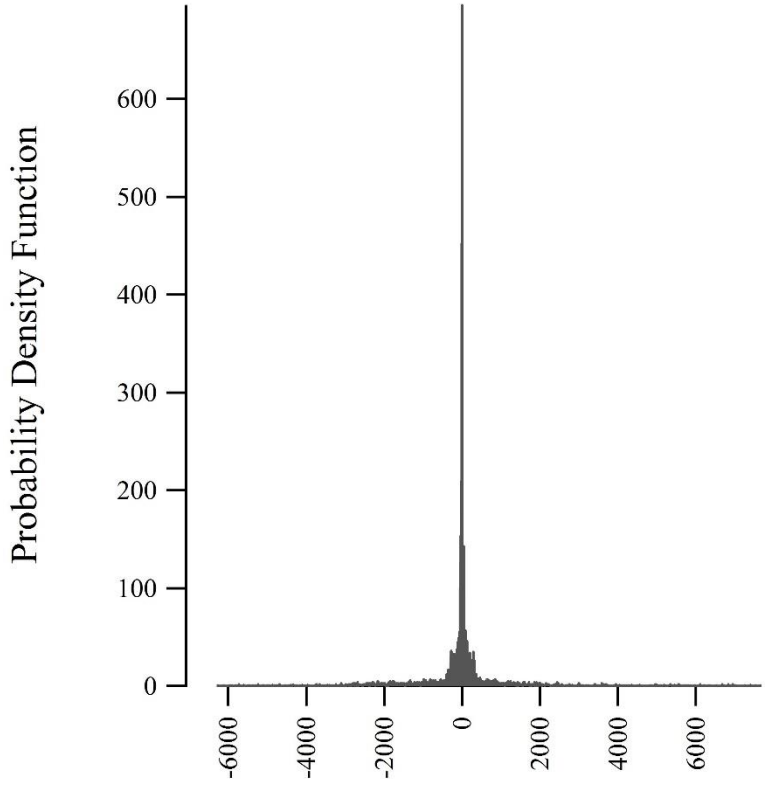
- Effect of water – rock interactions increasing from north to south, which is the flow direction
- Highlights the gaining behaviour of the river (increasing trend among surface water samples going downstream)

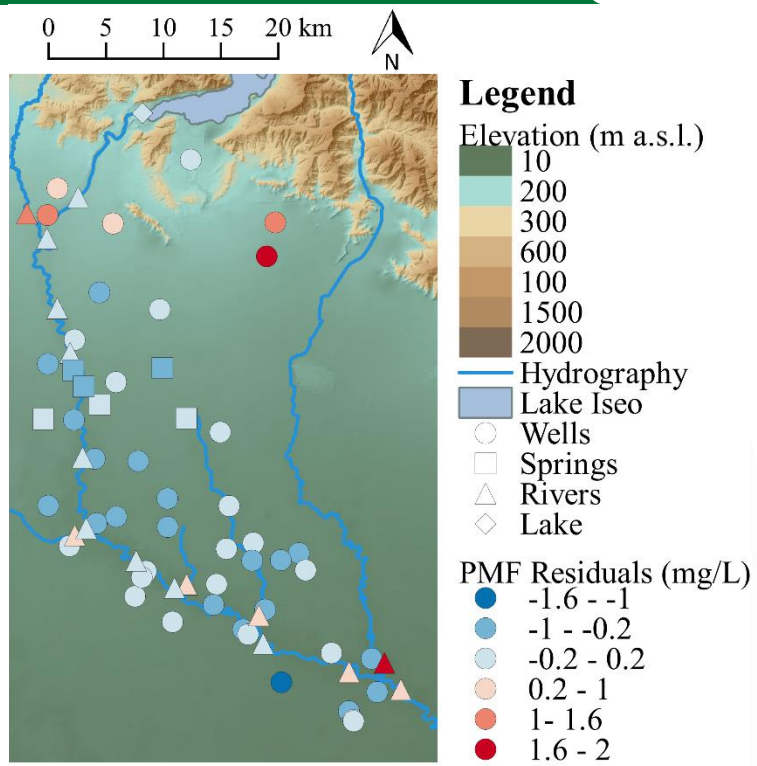
PMF Fac 5 – Anthropogenic impact



- Anthropogenic impact is higher in the higher plain (unconfined aquifer is more vulnerable and there is no reduction of the nitrate)
- This factor highlights the gaining behaviour of the river and the impact of tributaries

PMF Residual analysis



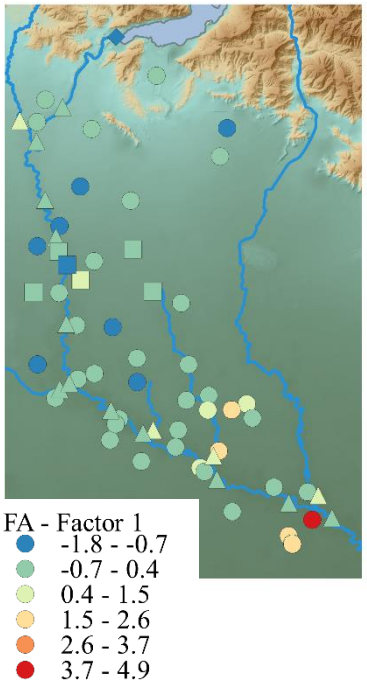


Residuals have a symmetric distribution

Residuals do not have a spatial pattern

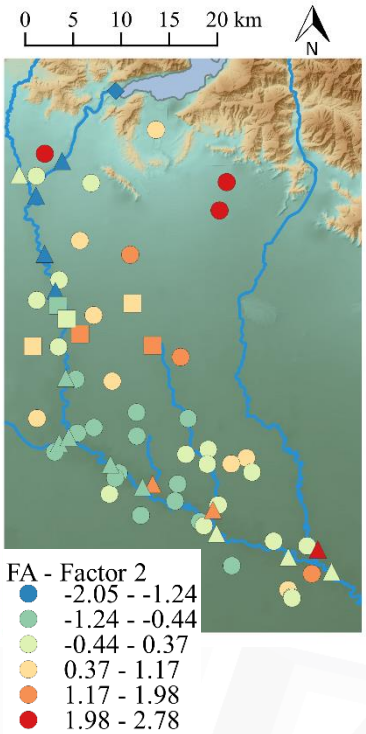
COMPARISON WITH FA

Factor analysis output consist in 3 factors:



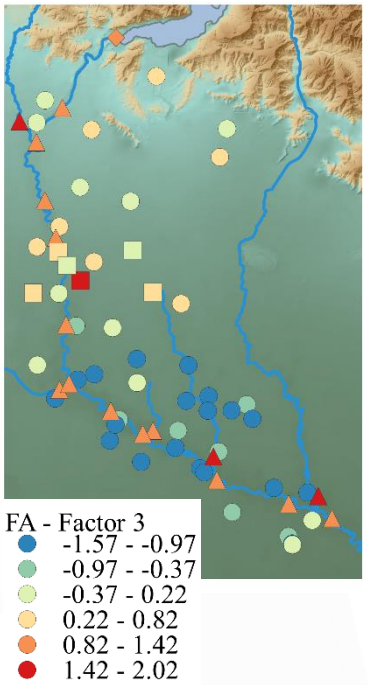
1- Reducing condition advanced + early stages

Ptot, NH_4, Na, As, Fe, Mn SO_4



2- Anthropogenic impact + water rock interactions

HCO_3, Cl, NO_3, Ca, Mg

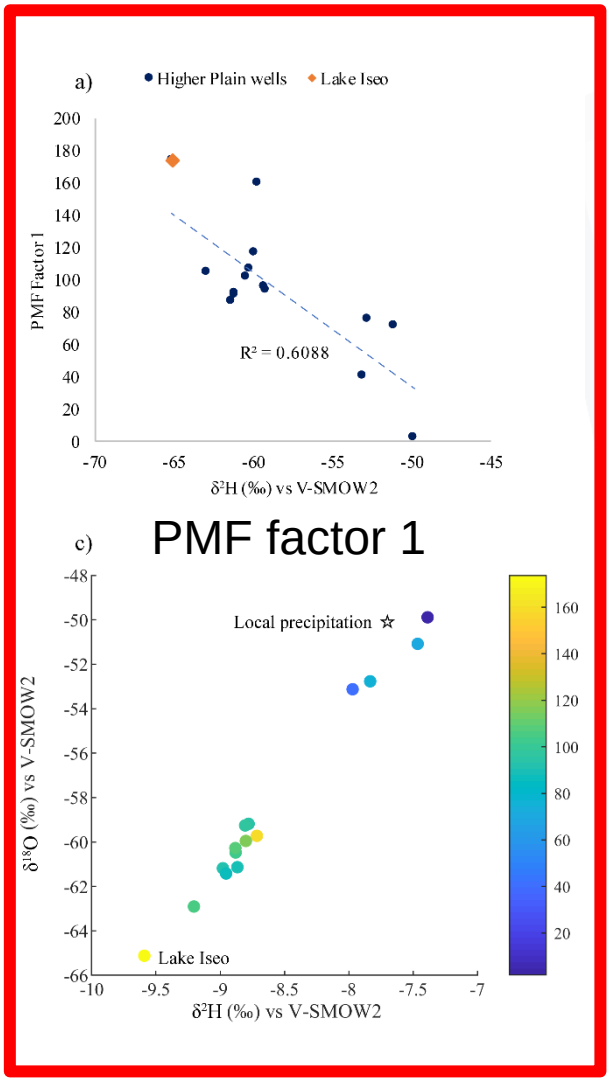


3- Surface water

O_2 , Cl, NO_3, SO_4, NH_4 K

COMPARISON WITH FA - validation

Validation of PMF factor 1 and FA factor 3 (surface water) with isotopic data



Two main sources of recharge in the higher plain aquifer:

- **Surface water from irrigation:** more negative values (<– precipitation from higher latitude and altitude)
- **Local precipitation:** less negative values (<– precipitation from local latitude and altitude)

Higher plain well samples more recharged by irrigation have isotopic signature closer to the signature of Lake Iseo

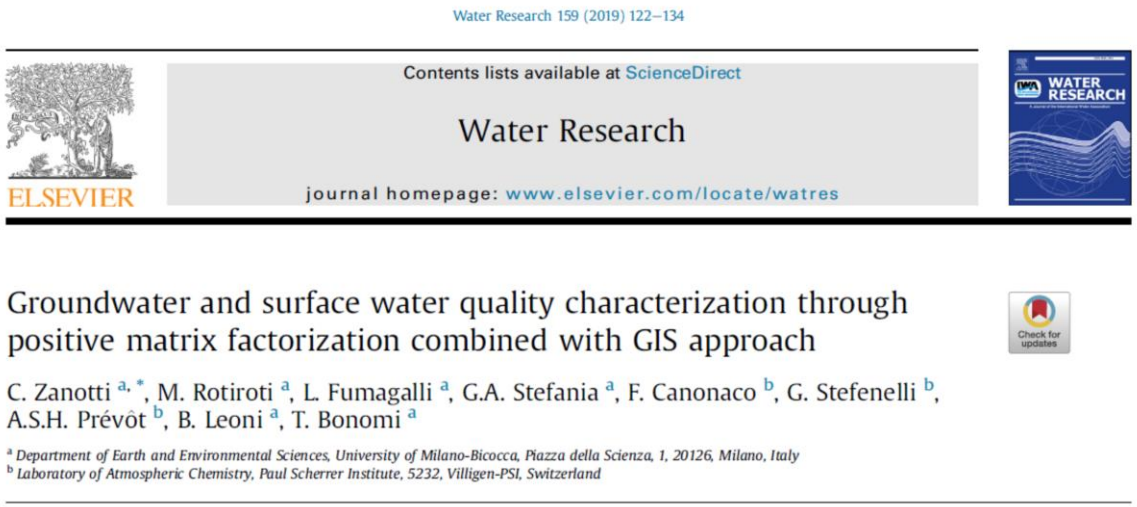
PMF fac 1 shows higher correlation with isotopic data

CONCLUSIONS

- Multivariate analysis allows for a synthetic description of the system
- All the main phenomena characterizing the system are identified and quantified
- PMF leads to a more environmentally interpretable solution as compared to FA

- PMF could be a suitable tool to perform exploratory analysis, supporting the development of the conceptual model

CONTACTS



Water Research 159 (2019) 122–134

Contents lists available at ScienceDirect

Water Research

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

Groundwater and surface water quality characterization through positive matrix factorization combined with GIS approach

C. Zanotti [a, *], M. Rotiroti [a], L. Fumagalli [a], G.A. Stefania [a], F. Canonaco [b], G. Stefenelli [b], A.S.H. Prévôt [b], B. Leoni [a], T. Bonomi [a]

[a] Department of Earth and Environmental Sciences, University of Milano-Bicocca, Piazza della Scienza, 1, 20126, Milano, Italy
[b] Laboratory of Atmospheric Chemistry, Paul Scherrer Institute, 5232, Villigen-PSI, Switzerland

For more information

Author contact: chiara.zanotti@unimib.it

full paper at: https://doi.org/10.1016/j.watres.2019.04.058