ECOLE MAROCAINE DES SCIENCES DE L'INGENIEUR
Membre de HONORIS UNITED UNIVERSITIES

Academic year 2025/2026

Project Report Data Analyst Hydraulic Structure

4th year Computer Engineering and Networks

31 december 2025

Presented by

AIT LAHCEN Achraf

Professor

Ms ELMKHALET Mouna

Table of Contents

Introduction: Presentation and explanation of the problem

Hydraulic structures play a central role in the sustainable management of water resources, particularly in regions subject to severe climatic constraints and increasing pressure from agricultural, industrial, and domestic uses. In this context, a thorough understanding of the hydraulic behavior of canals whether irrigation networks, diversions, or supply canals is essential for optimizing their use, limiting water losses, and preventing structural malfunctions.

The dataset studied describes a set of channels using eight physical and environmental variables: flow rate (m^3/s), water velocity (m/s), channel width and depth (m), hydraulic roughness, longitudinal slope (%), water temperature (°C), and siltation rate (%). These parameters simultaneously determine the transport capacity, morphological stability, and hydraulic efficiency of the structures. Poor control of these variables often results in siltation, excessive head loss, reduced available flow, and increased risks of erosion and structural degradation.

The aim of this work is therefore to analyze and characterize the hydraulic functioning of canals based on these variables, in order to identify typical operating profiles, the factors that most strongly control flow, and potentially critical situations (heavily silted canals, undersized cross-sections, unfavorable slopes, etc.). The ultimate goal is to develop decision-support tools for the diagnosis, classification, and improvement of hydraulic structures, using appropriate statistical and data analysis methods (factor analysis, classification, predictive models).

Chapter I: Presentation of the topic in the Moroccan context and application (PCA)

Morocco is characterized by significant spatial and temporal variability in water resources, exacerbated by recurring droughts and the effects of climate change. In this context, national policies have historically focused on water mobilization and development through a substantial network of dams and irrigation canals, particularly in the large areas managed by the Regional Offices for Agricultural Development (ORMVA). The performance of these infrastructure projects directly impacts the country's water security, agricultural productivity, and the resilience of rural areas.

However, many canals in the field suffer from recurring problems: high line losses, significant sediment deposits, clogging of sections, bank degradation, malfunctions of control mechanisms, and maintenance difficulties. These phenomena are strongly linked to local hydraulic characteristics: insufficient or poorly distributed flows, excessively low velocities promoting siltation, unsuitable gradients, high roughness due to the condition of the lining or deposits, as well as climatic constraints (water temperature, sediment-laden flood events).

In this Moroccan context, a detailed study of a dataset describing canals by flow rate, velocity, width, depth, roughness, slope, water temperature, and siltation becomes crucial. It allows us to:

- Better understand the relationships between these parameters under real-world operating conditions of Moroccan infrastructure.
- Identify canal types (sections in good condition, heavily silted canals, steep or undersized sections) that can guide maintenance and rehabilitation priorities.
- Provide quantitative data to support national strategies for modernizing irrigation networks (resurfacing, recalibration, sediment management, and flow control).

Thus, this chapter places the subject within the current challenges of water management in Morocco, where the optimization of the hydraulic functioning of canals becomes a major lever to secure irrigation, save the resource and extend the lifespan of existing infrastructure.

Part 1:

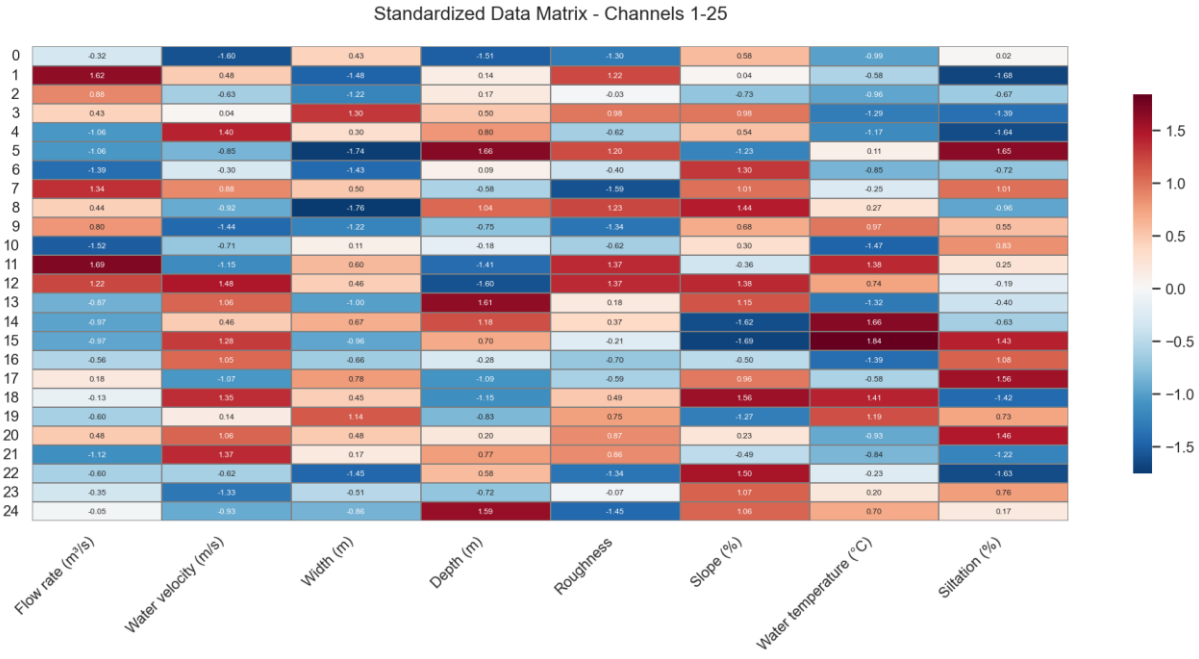
Mean and Standard Deviation by Criterion:

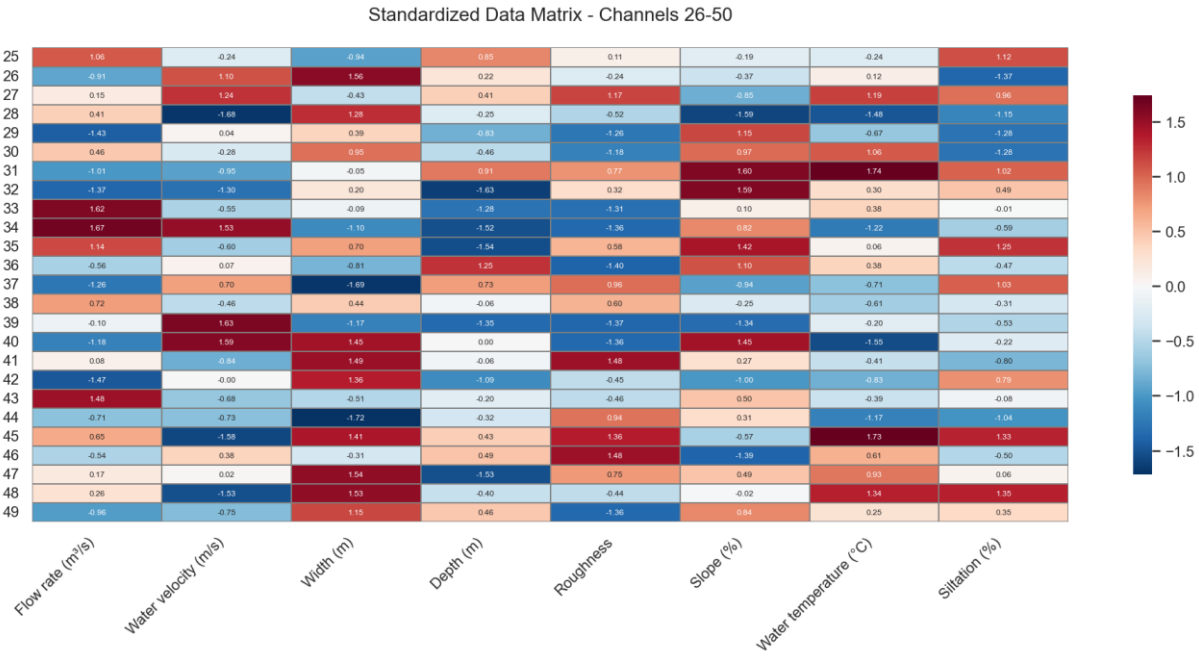
- Flow Rate (m^3/s): Mean = 94.04, Standard Deviation = 59.50
  → The average flow rate of the hydraulic structures is approximately 94 m^3/s, indicating a significant flow capacity. The high standard deviation (59.50) highlights considerable variability between sites, with some exhibiting flow rates well above or below the average.
- Water Velocity (m/s): Mean = 2.49, Standard Deviation = 1.47
  → The average water velocity is 2.5 m/s, a typical value for moderately fast currents. The standard deviation of 1.47 reflects significant dispersion, suggesting differences in hydraulic regime depending on the morphology and slope of the structures.
- Width (m): Mean = 26.36, Standard deviation = 14.38
  → The average width of the channel is approximately 26 m, indicating fairly large sections. The standard deviation of 14.38 shows significant heterogeneity among the structures, ranging from small canals to wide river sections.
- Depth (m): Mean = 5.17, Standard deviation = 2.79
  → An average depth of 5.2 m indicates relatively deep watercourses or reservoirs. The standard deviation of 2.79 highlights variability related to differences in configuration and water level.
- Roughness: Mean = 0.03, Standard deviation = 0.02
  → An average roughness of 0.03 reflects relatively regular and low-resistance flow conditions (smooth bottoms). The standard deviation (0.02) indicates significant heterogeneity between the different sections studied.
- Slope (%): Mean = 2.63, Standard deviation = 1.49
  → The average slope of 2.63% reflects moderately inclined flows, promoting a suitable water velocity. The standard deviation of 1.49 indicates a diversity of topographic conditions across the sites.
- Water temperature (°C): Mean = 15.79, Standard deviation = 6.68

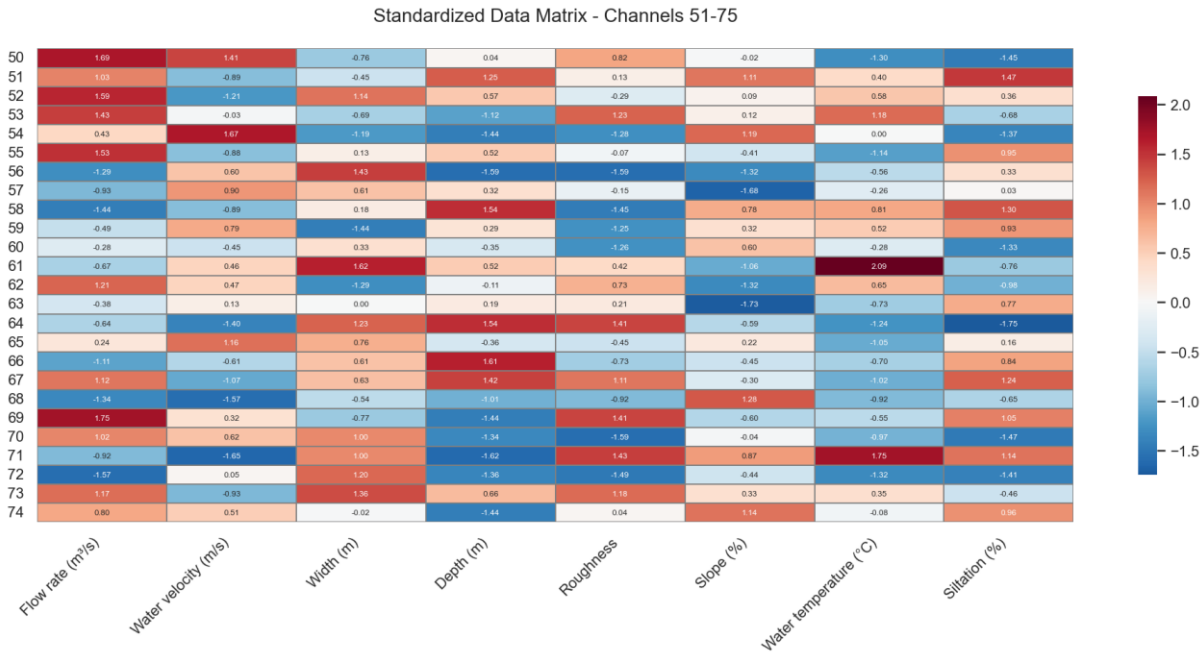
→ The average water temperature is approximately 15.8°C, indicating a temperate environment. The significant standard deviation (6.68) reveals substantial seasonal or geographic variability, between warmer and colder areas.

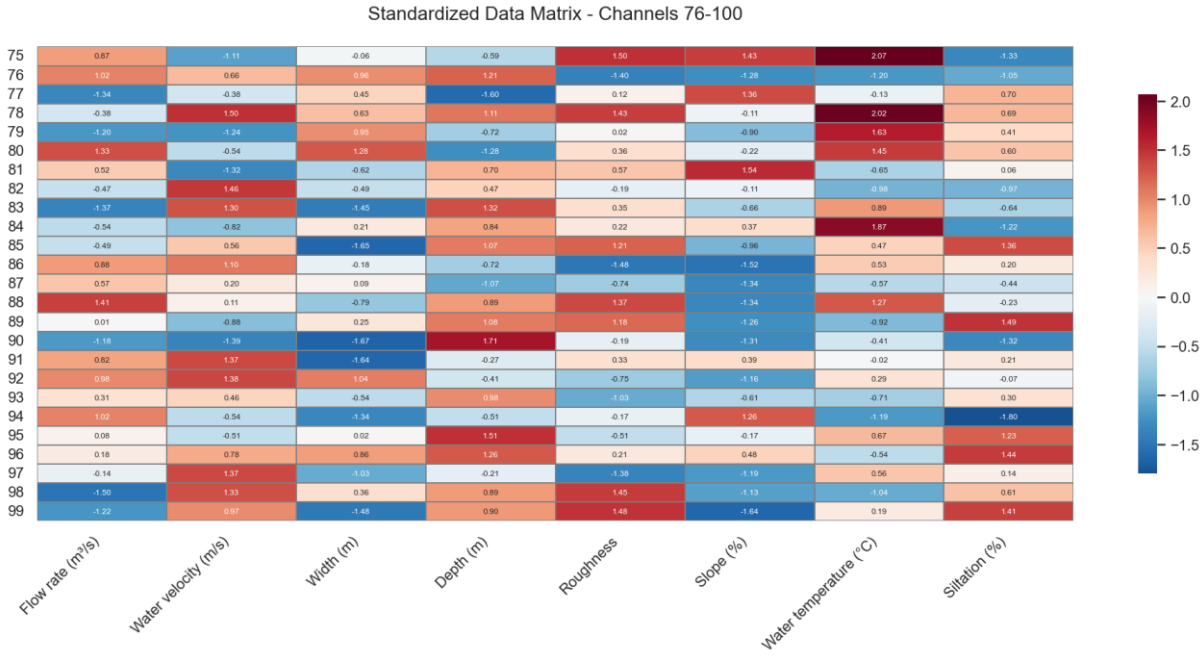
- Siltation (%): Mean = 52.60, Standard deviation = 28.34
  → With an average siltation of 52.6%, the structures exhibit a high level of sediment accumulation. The large standard deviation (28.34) indicates a significant disparity between sites, some being heavily silted while others remain more exposed.

Standardized Data Matrix









The standardized matrix shows, for each structure, how far each variable deviates from the mean in standard deviation units. Red cells correspond to values significantly above the mean (positive z-score), while blue cells indicate significantly below values (negative z-score).

General Interpretation

- Each row represents a structure (or section) and each column a variable: flow rate, velocity, width, depth, roughness, slope, temperature, and siltation.
- The values in the cells are the standardized centered scores (SCRs): around 0, the structures are close to the average; above 1 or below -1, they deviate significantly from the average behavior.

Identifying Extreme Structures

- Rows with a lot of red indicate structures that are more "extreme" than average: high flow rates, high velocities, large widths or depths, etc.
- Conversely, rows dominated by blue correspond to structures that are undersized or not in demand: low flow rates, slow velocities, narrow widths, low siltation, etc.

Behavior by Variable

- For flow rate and velocity, a marked alternation of positive and negative values is observed, confirming the high hydraulic variability already seen in the descriptive statistics (high standard deviations).
- Width and depth also show contrasts, reflecting very different dimensions from one structure to another (narrow channels vs. very wide and deep channels).

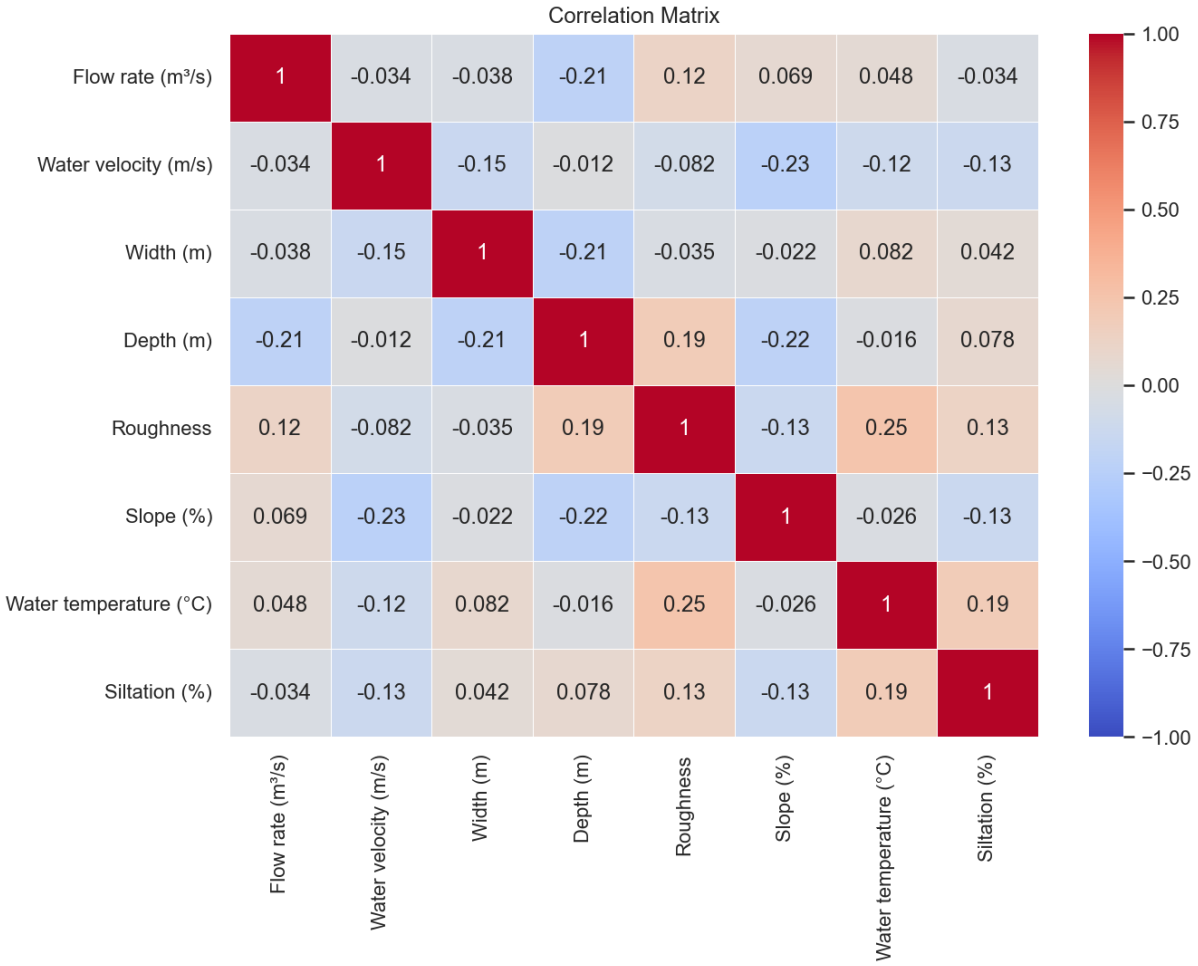
Roughness, Slope, and Temperature

- Roughness and slope fluctuate less extremely than flow rate or width, suggesting a relative homogeneity of bottom and topographic conditions, despite some very rough or very steep sections visible in red.
- Water temperature is shown in red and blue bands depending on the structure, reflecting seasonal or geographical differences (colder water at higher altitudes or during winter, warmer water elsewhere).

Siltation

- The siltation column alternates between very positive and very negative values, confirming strong sedimentary heterogeneity: some structures are heavily silted up (dark red), while others remain relatively clear (blue).
- This variability in siltation, combined with differences in flow rate and slope, indicates highly contrasting sediment transport conditions from one site to another.

Correlation Matrix



Flow Rate (m³/s)

- Water Velocity: -0.03 — Nearly no correlation. Flow rate does not systematically vary with local water velocity, meaning that flow rate differences primarily originate from the wetted cross-section (width × depth) rather than the velocity itself.
- Width: -0.04 — Very weak and negative correlation. Wider structures do not necessarily have higher flow rates; a large channel may have low flow, and conversely, a narrower channel may carry a significant flow.
- Depth: -0.21 — Weak to moderate negative correlation. Deeper sections tend to be slightly associated with lower flow rates, which may reflect areas of storage or slowing (widening, reservoirs) rather than stretches of intense flow.
- Roughness: 0.12 — Weak positive correlation. A slight link exists between higher flow rate and greater roughness, possibly related to high-flow sections where the streambed is coarser (boulders, gravel).
- Slope: 0.07 — Very weak positive correlation. Slope only marginally explains flow rate variations, reinforcing the idea that flow rate is primarily determined by the catchment area and upstream inputs.
- Temperature: 0.05 — Nearly no correlation. There is no clear relationship between water temperature and flow rate across the entire sample.

- Nearly no correlation. The siltation rate does not appear to be directly related to the average flow rate, but rather to local sediment transport dynamics.

Water velocity (m/s)

- Flow rate: -0,03 — Nearly no correlation. A higher flow rate does not necessarily mean a higher velocity, because the geometry of the section compensates (width/depth).
- Width: -0,15 — Weak negative correlation. Wider channels tend to have slightly lower velocities, consistent with energy dissipation over a larger cross-section.
- Depth: -0,01 — Nearly no correlation. Depth alone does not explain velocity; other parameters such as slope and roughness play a major role.
- Roughness: -0.08 — Weak negative correlation. Higher roughness (coarser bottoms) is associated with slightly lower velocities, reflecting increased head loss.
- Slope: -0.23 — Weak to moderate negative correlation, somewhat counterintuitive. In this dataset, the steepest sections appear to be associated with lower velocities, which may be explained by rougher or more confined channels where energy is dissipated.
- Temperature: -0.12 — Weak negative correlation. Colder waters are slightly associated with higher velocities, or vice versa; this may reflect seasonal or contextual differences (mountain rivers).
- Siltation: -0.13 — Weak negative correlation. Heavily silted sections tend to have lower velocities, which makes sense: slow flow promotes the deposition of fine sediments.

Width (m)

- Flow rate: -0,04 — Very weak negative correlation. A large width does not necessarily imply a high discharge; some oversized channels may have low flow rates.
- Vitesse: -0,15 — Weak negative correlation. Wider streams tend to have lower velocities because the water is spread over a larger cross-section.
- Depth: -0.21 — Weak to moderate negative correlation. In our data, very wide channels are often shallower, reflecting a geometric trade-off between width and depth.
- Roughness: -0.04 — Very weak negative correlation. Width is virtually unrelated to bottom roughness.
- Slope: -0.02 — Nearly no correlation. Variations in slope are not systematically accompanied by a change in width.
- Temperature: 0.08 — Weak positive correlation. Wider sections are slightly associated with slightly warmer water, possibly because they are more exposed to solar radiation.
- Siltation: 0.04 — Very weak positive correlation. Wider channels tend to be slightly more silted, but the relationship remains very limited.

Depth (m)

- Flow rate: -0.21 — Weak to moderate negative correlation. The deepest sections of our sample are not necessarily the highest discharge areas; they may be areas of slowing or storage (plunges, pools, reservoirs).
- Velocity: -0.01 — Nearly no correlation. Depth alone does not determine flow velocity.
- Width: -0.21 — Weak to moderate negative correlation. Deeper sections tend to be narrower, corresponding to V-shaped or steep channel profiles.
- Roughness: 0.19 — Weak to moderate positive correlation. Deeper areas are slightly associated with higher roughness, for example, boulder or rocky channels.
- Slope: -0.22 — Weak to moderate negative correlation. Very steep sections are generally shallower, as in mountain streams.
- Temperature: -0.02 — Near-zero correlation. Depth does not significantly affect the average water temperature in this dataset.
- Siltation: 0.08 — Weak positive correlation. Deeper areas retain slightly more fine sediment, which is consistent with lower local velocities in the valley bottom.

Roughness

- Flow rate: 0.12 — Weak positive correlation. Higher discharge reaches tend to have a slightly rougher channel (coarse materials), characteristic of rivers active in sediment transport.
- Velocity: -0.08 — Weak negative correlation. Greater roughness slightly reduces average velocity, corresponding to head losses due to friction.
- Width: -0.04 — Very weak negative correlation. Roughness is almost independent of channel width.
- Depth: 0.19 — Weak to moderate positive correlation. Deeper sections often have a coarser channel, for example channels incised into the substrate.
- Slope: -0.13 — Weak negative correlation. In our data, steeper slopes are not associated with greater roughness; this may indicate modifications (plaster, concrete, smoothed riprap).
- Temperature: 0.25 — Moderate positive correlation. Rougher sections correspond to slightly warmer water, possibly related to slow velocity stretches that promote heating.
- Siltation: 0.13 — Weak positive correlation. A rougher channel retains slightly more fine sediment in its irregularities, thus increasing the observed siltation.

Slope (%)

- Flow rate: 0.07 — Very weak positive correlation. Discharge is almost entirely unaffected by slope in this sample, indicating a strong influence from the catchment area and upstream inputs.

- Velocity: -0.23 — Weak to moderate negative correlation. Contrary to the theoretical case, velocity decreases as the slope increases, probably because steeper sections are rougher, more incised, or exhibit unique losses (weirs, structures).
- Width: -0.02 — Nearly no correlation. Slope is not related to channel width.
- Depth: -0.22 — Weak to moderate negative correlation. Steep slopes are often associated with shallower channels, typical of torrential streams.
- Roughness: -0.13 — Weak negative correlation. The steeper sections in our dataset appear to be more engineered or less rough, which may reflect the presence of artificial structures.
- Water Temperature: -0.03 — Near-zero correlation. Slope has no significant effect on average water temperature.
- Siltation: -0.13 — Weak negative correlation. Steeper sections have less siltation, consistent with higher local velocities that remobilize sediments.

Water temperature (°C)

- Flow Rate: 0.05 — Nearly no correlation. Flow rate does not directly control temperature; the main influence comes from climate and water source.
- Velocity: -0.12 — Weak negative correlation. Slower-moving waters tend to be slightly warmer because they remain exposed to solar radiation for longer.
- Width: 0.08 — Weak positive correlation. Wider channels, often better exposed, have slightly higher temperatures.
- Depth: -0.02 — Nearly no correlation. Average depth does not significantly affect overall temperature in our data.
- Roughness: 0.25 — Moderate positive correlation. Rougher sections, which are more dissipative and often have slower flow, are associated with slightly warmer waters.
- Slope: -0.03 — Nearly no correlation. Slope does not directly affect temperature.
- Siltation: 0.19 — Weak to moderate positive correlation. Areas with high siltation, less well mixed and sometimes more stagnant soils tend to warm up more.

Siltation (%)

- Flow rate: -0.03 — Nearly no correlation. The extent of siltation does not depend directly on the average flow rate, but rather on local deposition and erosion conditions.
- Velocity: -0.13 — Weak negative correlation. Low velocities promote the deposition of fine particles, resulting in greater siltation in slow-moving sections.
- Width: 0.04 — Very weak positive correlation. Wider sections show slightly higher siltation, but the link remains tenuous.
- Depth: 0.08 — Weak positive correlation. Deeper areas act as traps for fine sediments, increasing siltation.

- Roughness: 0.13 — Weak positive correlation. A rougher channel retains more particles in its irregularities.
- Slope: -0.13 — Weak negative correlation. Steep slopes are less silted up because the available energy allows for the remobilization of deposits.
- Temperature: 0.19 — Weak to moderate positive correlation. Warmer waters, often more stagnant or located in lowland areas, are associated with higher siltation rates.

Eigenvalues and explained variance

Component 1: 1.5686 | Explained inertia: 0.1941 (19.41%) - Cumulative: 19.41%

Component 2: 1.4662 | Explained inertia: 0.1814 (18.14%) - Cumulative: 37.56%

Component 3: 1.1356 | Explained inertia: 0.1405 (14.05%) - Cumulative: 51.61%

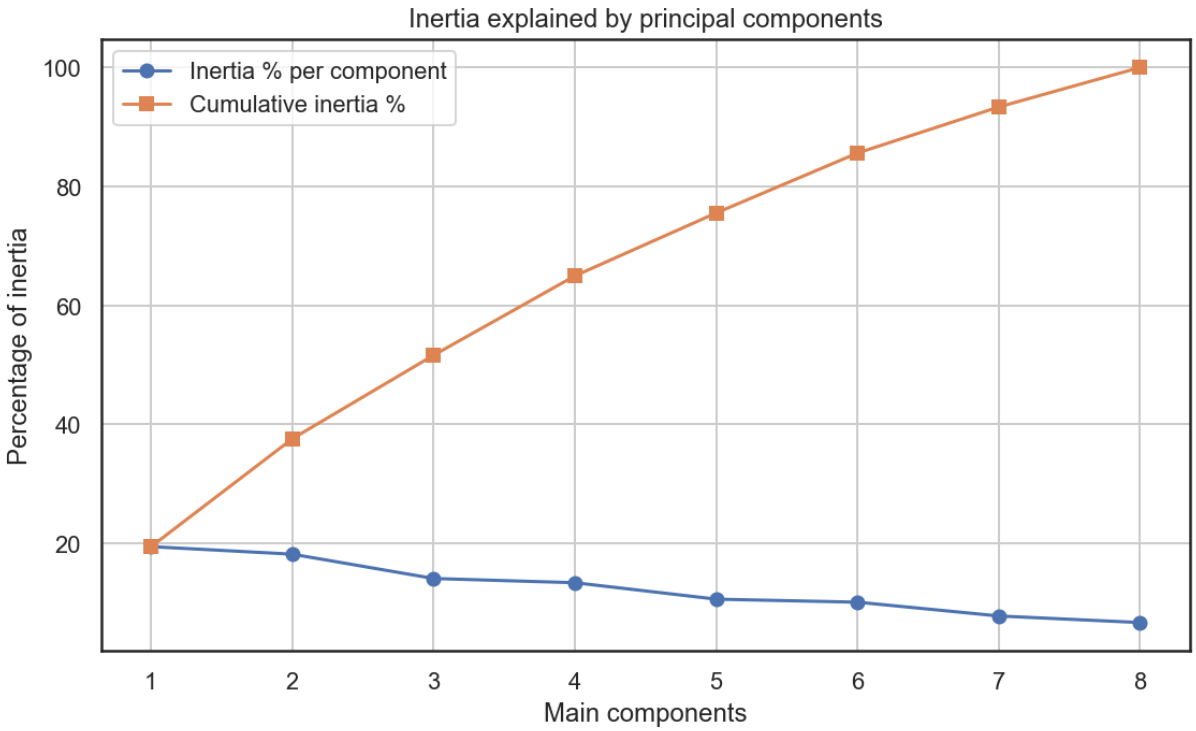
Component 4: 1.0782 | Explained inertia: 0.1334 (13.34%) - Cumulative: 64.95%

Component 5: 0.8548 | Explained inertia: 0.1058 (10.58%) - Cumulative: 75.53%

Component 6: 0.8140 | Explained inertia: 0.1007 (10.07%) - Cumulative: 85.60%

Component 7: 0.6263 | Explained inertia: 0.0775 (7.75%) - Cumulative: 93.35%

Component 8: 0.5372 | Explained inertia: 0.0665 (6.65%) - Cumulative: 100.00%



Interpretation of Principal Components

- Component 1: 1.5686 | 19.41% – Cumulative: 19.41%

  This first component captures the largest share of variability, but it only summarizes approximately 19.4% of the total information. This means that a single dimension is

insufficient to describe the structure of our 8 variables: the phenomenon studied is clearly multifactorial and cannot be reduced to a single dominant axis.

- Component 2: 1.4662 | 18.14% – Cumulative: 37.56%

  The second component adds a significant amount of information (18.1%), bringing the explained variance to nearly 37.6%. The first two axes thus characterize slightly more than a third of the data structure, allowing for an initial visualization (factorial plane 1–2) but with a still considerable loss of information.

- Component 3: 1.1356 | 14.05% – Cumulative: 51.61%

  The third component contributes another 14.1% of the variance. With three axes, PCA recovers slightly more than half of the information (51.6%), demonstrating that several independent dimensions contribute to the variability of hydraulic characteristics.

- Component 4: 1.0782 | 13.34% – Cumulative: 64.95%

  The fourth component explains an additional 13.3%, bringing the cumulative total to approximately 65%. The first four axes thus concentrate on the two levels of total variance: they represent a good compromise for the overall interpretation, while accepting a loss of about 35% of information.

- Component 5: 0.8548 | 10.58% – Cumulative: 75.53%

  The fifth component, although its eigenvalue is less than 1, still retains 10.6% of the variance. Including it increases the information yield to 75.5%, useful if you want to refine the analysis without excessively multiplying the dimensions.

- Component 6: 0.8140 | 10.07% – Cumulative: 85.60%

  The sixth component adds approximately 10%, bringing the explained variance to 85.6%. At this level, the essential structure of the data is captured, but at the cost of a model that is already difficult to visualize (6 dimensions).

- Component 7: 0.6263 | 7.75% – Cumulative: 93.35%

  The seventh component provides only a moderate gain (7.8%). It highlights finer nuances or special cases in the scatter plot, interesting for detailed analysis but less so for a simple summary.

- Component 8: 0.5372 | 6.65% – Cumulative: 100%

  The last component makes up the remaining 6.6%, so that all eight axes together account for 100% of the variance. It mainly corresponds to noise or very specific contrasts, which contribute only marginally to overall understanding.

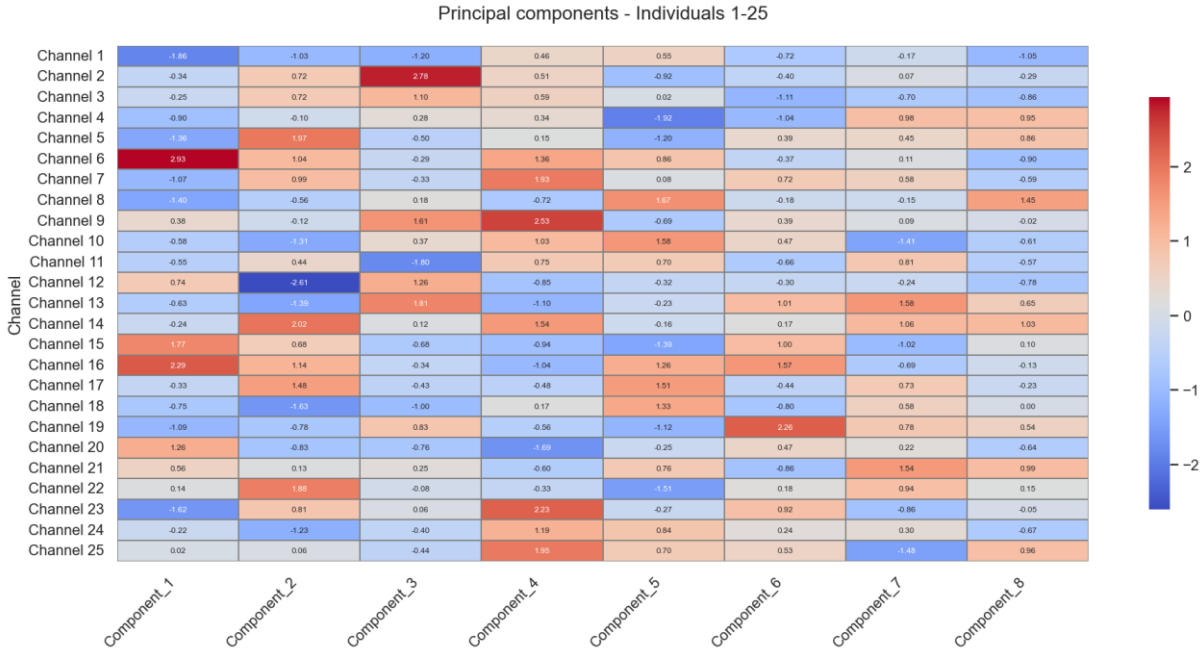
Percentage of information recovered / lost

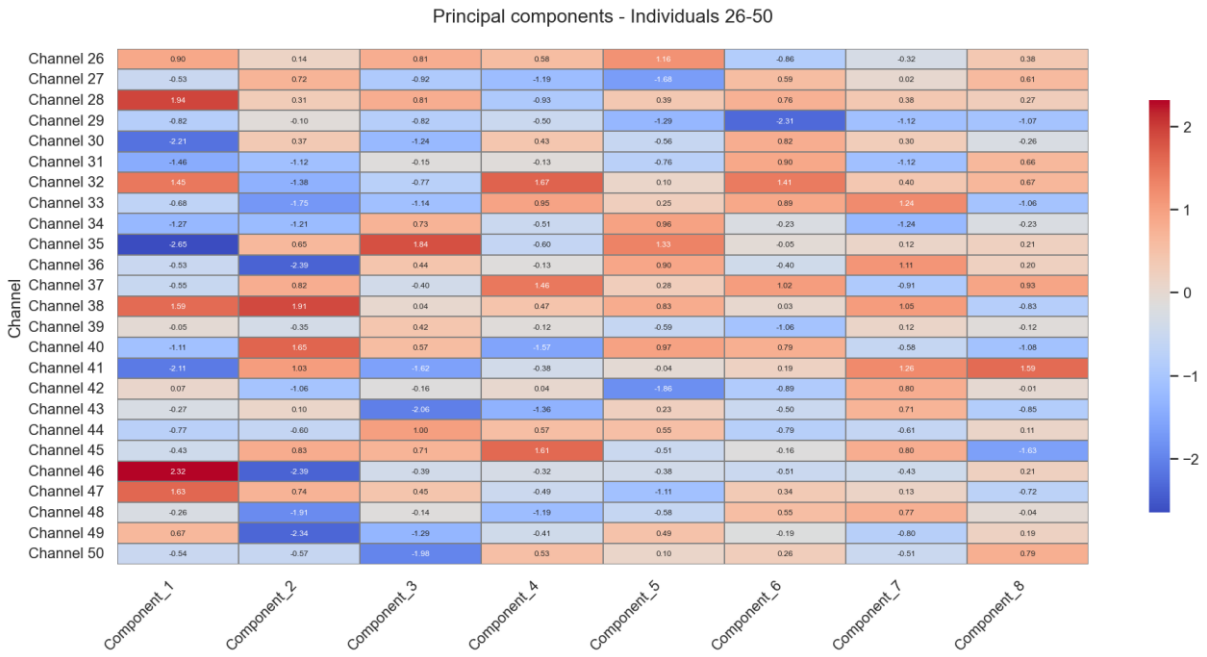
- With 2 components:
  - Information recovered: 37.56%
  - Information lost: 62.44%

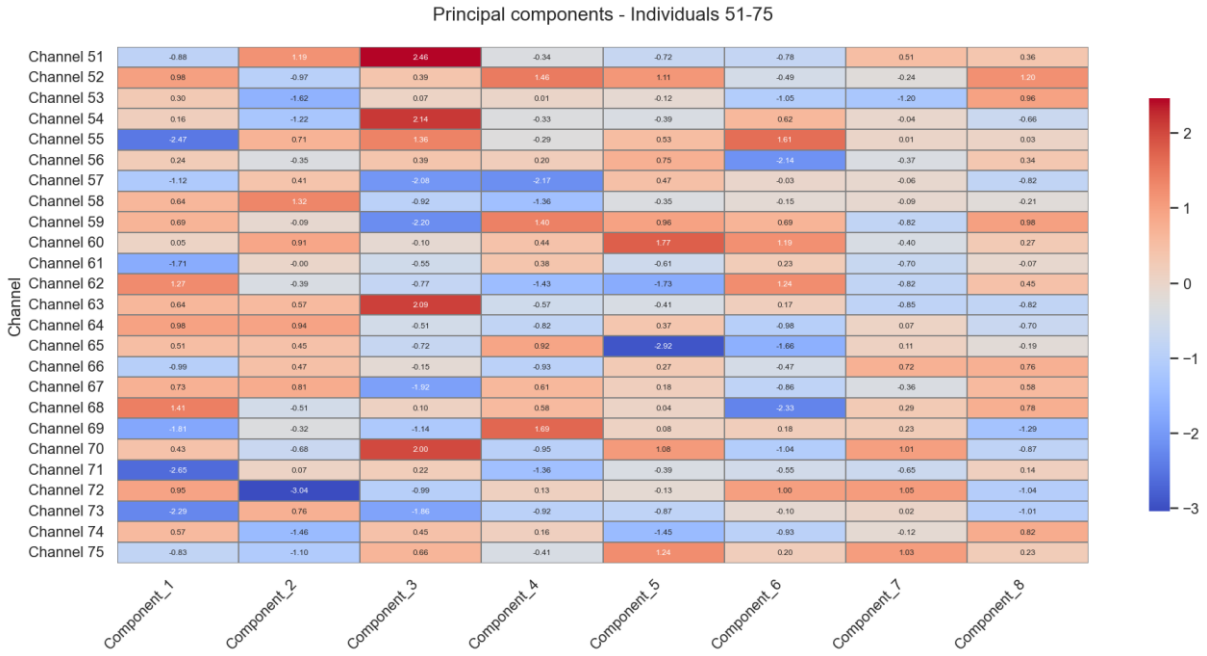
- With 3 components:
    - Information recovered: 51.61%
    - Information lost: 48.39%
- With 4 components:
    - Information recovered: 64.95%
    - Information lost: 35.05%
- With all components (8):
    - Information recovered: 100%
    - Information lost: 0%

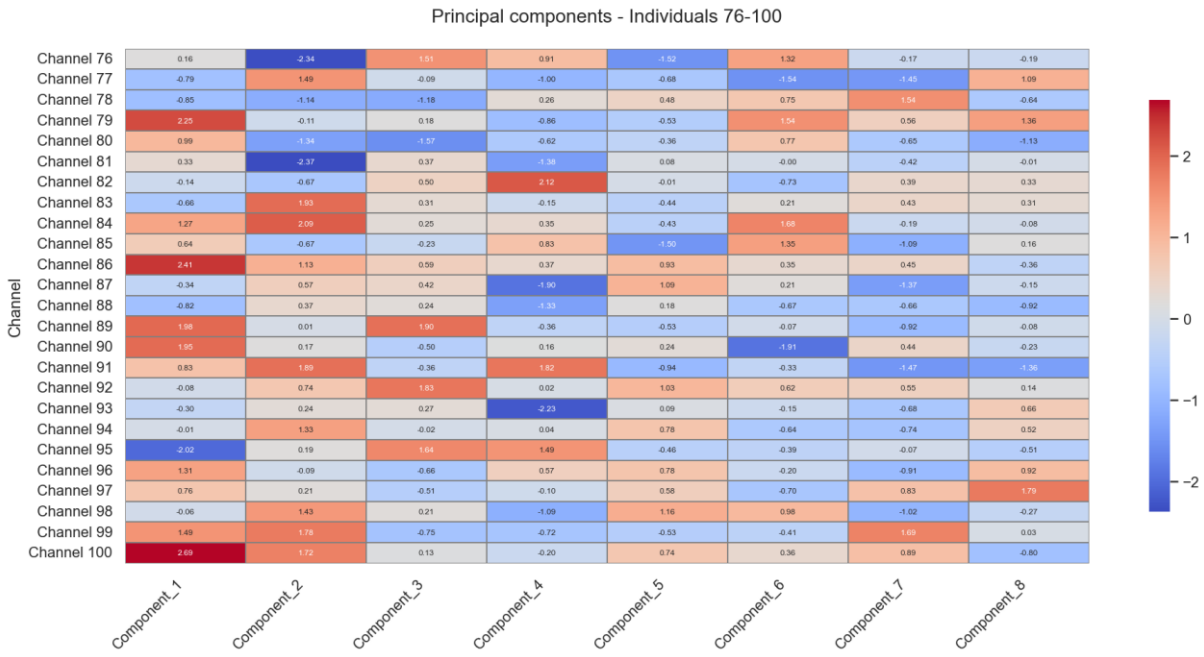
Part 2:

Principal components of the individuals and variables




Principal components - Individuals 26-50
Channel
Channel 26 0.90 0.14 0.81 0.58 1.16 -0.86 -0.32 0.38
Channel 27 -0.53 0.72 -0.92 -1.19 -1.68 0.59 0.02 0.61
Channel 28 1.94 0.31 0.81 -0.93 0.39 0.76 0.38 0.27
Channel 29 -0.82 -0.10 -0.82 -0.50 -1.29 -2.31 -1.12 -1.07
Channel 30 -2.21 0.37 -1.24 0.43 -0.56 0.82 0.30 -0.26
Channel 31 -1.46 -1.12 -0.15 -0.13 -0.76 0.90 -1.12 0.66
Channel 32 1.45 -1.38 -0.77 1.67 0.10 1.41 0.40 0.67
Channel 33 -0.68 -1.75 -1.14 0.95 0.25 0.89 1.24 -1.06
Channel 34 -1.27 -1.21 0.73 -0.51 0.96 -0.23 -1.24 -0.23
Channel 35 -2.65 0.65 1.84 -0.60 1.33 -0.05 0.12 0.21
Channel 36 -0.53 -2.39 0.44 -0.13 0.90 -0.40 1.11 0.20
Channel 37 -0.55 0.82 -0.40 1.46 0.28 1.02 -0.91 0.93
Channel 38 1.59 1.91 0.04 0.47 0.83 0.03 1.05 -0.83
Channel 39 -0.05 -0.35 0.42 -0.12 -0.59 -1.06 0.12 -0.12
Channel 40 -1.11 1.65 0.57 -1.57 0.97 0.79 -0.58 -1.08
Channel 41 -2.11 1.03 -1.62 -0.38 -0.04 0.19 1.26 1.59
Channel 42 0.07 -1.06 -0.16 0.04 -1.86 -0.89 0.80 -0.01
Channel 43 -0.27 0.10 -2.06 -1.36 0.23 -0.50 0.71 -0.85
Channel 44 -0.77 -0.60 1.00 0.57 0.55 -0.79 -0.61 0.11
Channel 45 -0.43 0.83 0.71 1.61 -0.51 -0.16 0.80 -1.63
Channel 46 2.32 -2.39 -0.39 -0.32 -0.38 -0.51 -0.43 0.21
Channel 47 1.63 0.74 0.45 -0.49 -1.11 0.34 0.13 -0.72
Channel 48 -0.26 -1.91 -0.14 -1.19 -0.58 0.55 0.77 -0.04
Channel 49 0.67 -2.34 -1.29 -0.41 0.49 -0.19 -0.80 0.19
Channel 50 -0.54 -0.57 -1.98 0.53 0.10 0.26 -0.51 0.79
Component_1
Component_2
Component_3
Component_4
Component_5
Component_6
Component_7
Component_8
2
1
0
−1
−2


Principal components - Individuals 51-75
Channel
Channel 51 -0.88 1.19 2.46 -0.34 -0.72 -0.78 0.51 0.36
Channel 52 0.98 -0.97 0.39 1.46 1.11 -0.49 -0.24 1.20
Channel 53 0.30 -1.62 0.07 0.01 -0.12 -1.05 -1.20 0.96
Channel 54 0.16 -1.22 2.14 -0.33 -0.39 0.62 -0.04 -0.66
Channel 55 -2.47 0.71 1.36 -0.29 0.53 1.61 0.01 0.03
Channel 56 0.24 -0.35 0.39 0.20 0.75 -2.14 -0.37 0.34
Channel 57 -1.12 0.41 -2.08 -2.17 0.47 -0.03 -0.06 -0.82
Channel 58 0.64 1.32 -0.92 -1.36 -0.35 -0.15 -0.09 -0.21
Channel 59 0.69 -0.09 -2.20 1.40 0.96 0.69 -0.82 0.98
Channel 60 0.05 0.91 -0.10 0.44 1.77 1.19 -0.40 0.27
Channel 61 -1.71 -0.00 -0.55 0.38 -0.61 0.23 -0.70 -0.07
Channel 62 1.27 -0.39 -0.77 -1.43 -1.73 1.24 -0.82 0.45
Channel 63 0.64 0.57 2.09 -0.57 -0.41 0.17 -0.85 -0.82
Channel 64 0.98 0.94 -0.51 -0.82 0.37 -0.98 0.07 -0.70
Channel 65 0.51 0.45 -0.72 0.92 -2.92 -1.66 0.11 -0.19
Channel 66 -0.99 0.47 -0.15 -0.93 0.27 -0.47 0.72 0.76
Channel 67 0.73 0.81 -1.92 0.61 0.18 -0.86 -0.36 0.58
Channel 68 1.41 -0.51 0.10 0.58 0.04 -2.33 0.29 0.78
Channel 69 -1.81 -0.32 -1.14 1.69 0.08 0.18 0.23 -1.29
Channel 70 0.43 -0.68 2.00 -0.95 1.08 -1.04 1.01 -0.87
Channel 71 -2.65 0.07 0.22 -1.36 -0.39 -0.55 -0.65 0.14
Channel 72 0.95 -3.04 -0.99 0.13 -0.13 1.00 1.05 -1.04
Channel 73 -2.29 0.76 -1.86 -0.92 -0.87 -0.10 0.02 -1.01
Channel 74 0.57 -1.46 0.45 0.16 -1.45 -0.93 -0.12 0.82
Channel 75 -0.83 -1.10 0.66 -0.41 1.24 0.20 1.03 0.23
Component_1
Component_2
Component_3
Component_4
Component_5
Component_6
Component_7
Component_8
2
1
0
−1
−2
−3



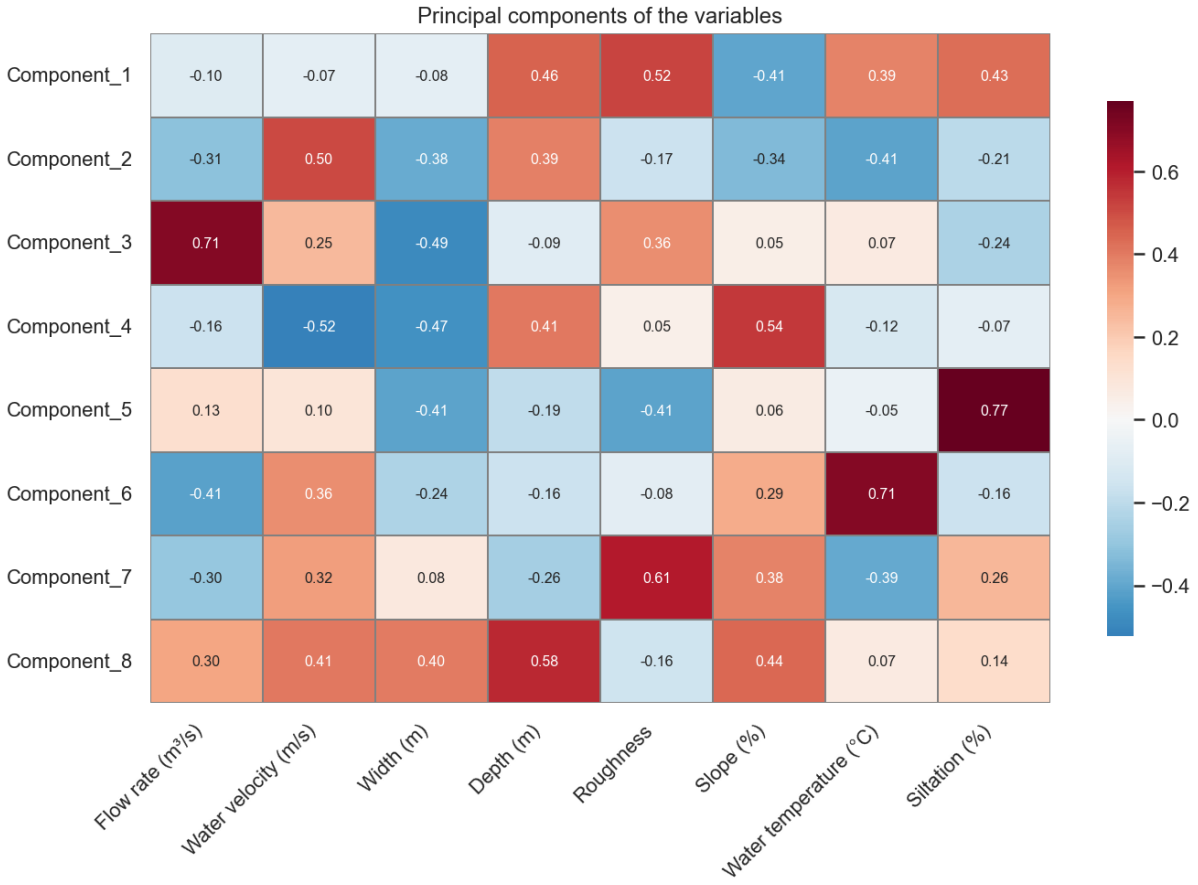
This heatmap shows, for each channel, its position on the 8 principal components (PCA scores), with high values in red and low values in blue.

General Interpretation

- Each row = one channel, each column = one principal component: a red value indicates a channel well above average on that component, a blue value a channel well below average.
- Channels whose lines alternate strongly between red and blue are atypical profiles (very pronounced on some axes, very weak on others), while light gray/lightly colored lines represent channels close to average behavior.

Role of the First Components

- On components 1 to 3 (the most significant in terms of inertia), some channels stand out clearly in red or blue: these are the ones that most structure the overall variability of the dataset and deserve detailed analysis (for example, to understand what types of geometry or hydraulic regime they represent).
- The following components (4 to 8) show finer contrasts: they distinguish subgroups of channels with specific behaviors (for example, very silted, very steep or particularly rough channels that the first axes did not clearly separate).



This heatmap shows the correlations between each principal component and the variables: strong values (red) indicate the variables that most strongly influence each axis.

Component 1

- Strongly correlated with roughness (0.52), then with depth, temperature, and siltation (coefficients around 0.4).
- This axis therefore contrasts channels with rough, deep, warm, and silted channels (positive scores) with smoother, shallower, and less silted channels (negative scores): it is a "channel condition/sedimentation" axis.

Component 2

- Positively influenced by velocity (0.50) and negatively by slope (−0.34) and, to a lesser extent, width.
- It differentiates channels with high velocity but moderate slope (probably wide, modified) from steeper channels with lower velocity, where roughness and losses dissipate energy.

Component 3

- Strongly and positively correlated with flow rate (0.71) and somewhat correlated with velocity, with a marked negative contribution from width (−0.49).

- This axis represents high-flow, relatively narrow channels versus wider but lower-flow channels, typical of the differences between concentrated sections and wide floodplains.

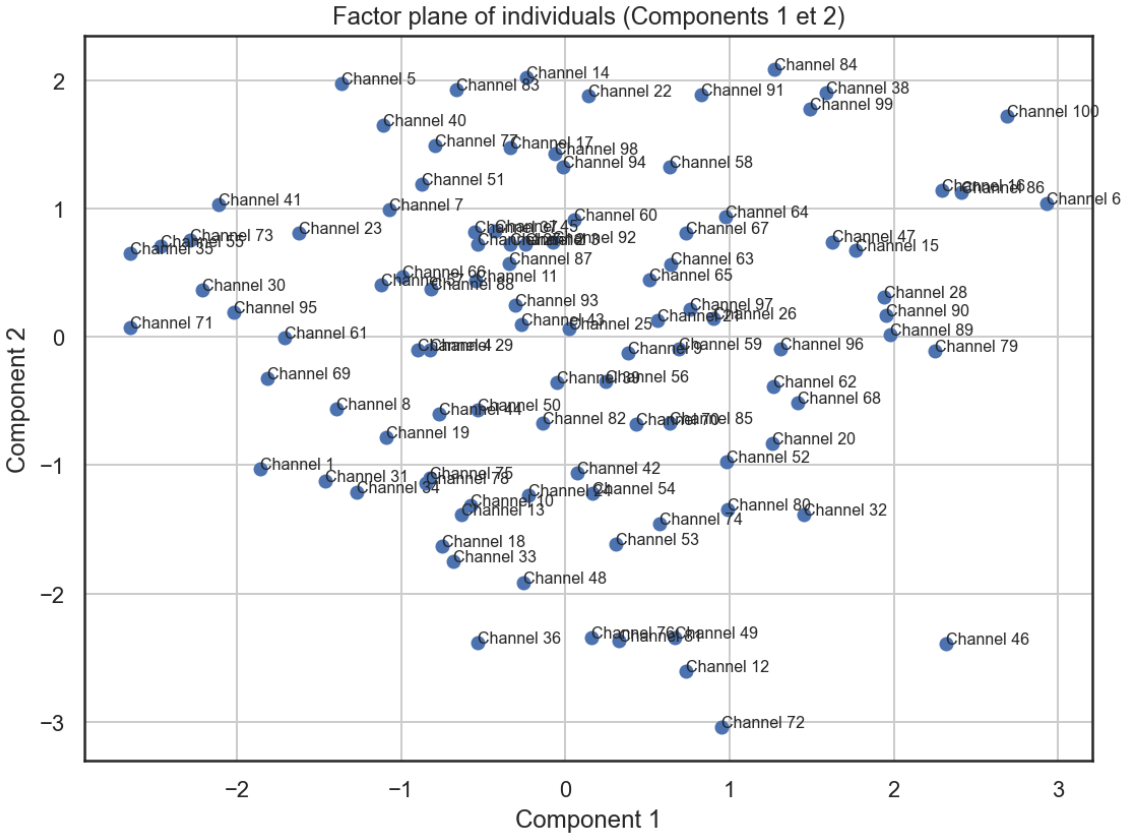
Component 4

- Positively correlated with depth (0.41) and especially with slope (0.54), negatively with velocity (−0.52) and width (−0.47).
- It characterizes steep, rough, relatively narrow, and slow-moving sections, as opposed to wider, less rough sections with higher velocity: a "torrential incised vs. wide-bore channel" axis.

Component 5 à 8 (axes plus fins)

- Component 5 is primarily driven by siltation (0.77) and, to a lesser extent, by width and slope: it isolates particularly clogged channels from the rest.
- Component 6 highlights temperature (0.71) and, to a lesser extent, velocity; it distinguishes warmer channels (often lowland or slowly renewed) from colder channels.
- Component 7 is related to roughness (0.61) and, to a lesser extent, velocity; it refines the separation between very rough and smoother channels within groups already similar along the first axes.
- Component 8 strongly contrasts depth (0.58), width, and velocity: it is a secondary axis describing differences in residual geometric dimensions between channels.

Principal component scores of individuals on the factor plane



This graph represents the factorial plane (axes 1 and 2) of the channels, that is, the projection of each channel onto the first two principal components.

Reading the axes

- The horizontal axis (Component 1) primarily reflects a roughness-depth-temperature-siltation gradient: towards the right, we find rougher, deeper, warmer, and siltier channels; towards the left, smoother, shallower, and less silted channels.
- The vertical axis (Component 2) contrasts channels with high velocity and a shallower slope (at the top) with steeper channels but lower velocity (at the bottom).

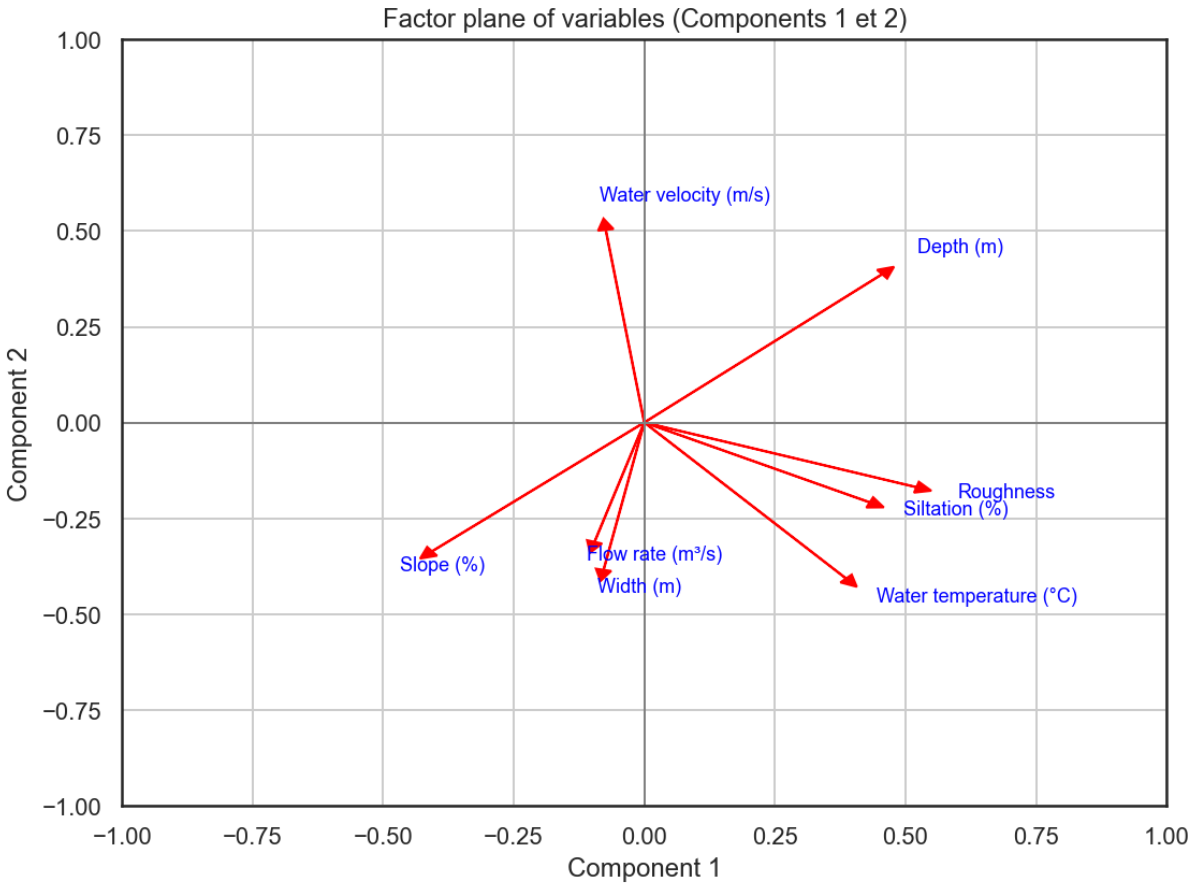
Channel position

- Channels near the center (around 0,0) exhibit average behavior on these two axes: no particularly pronounced characteristics in terms of channel/sedimentation or slope-velocity relationship.
- The channels on the far right (e.g., Channel 6, Channel 100, Channel 86) are rough, deep, and silted sections, typical of more silted-up or morphologically complex stretches.
- The channels on the far left (e.g., some numbered channels around -2 on axis 1) correspond to smoother, less silted channels, possibly more artificially modified or well-maintained.

- The channels at the top of the graph (e.g., Channel 14, Channel 83, Channel 38, Channel 84, Channel 99) have relatively high flow velocities for a moderate gradient, often associated with more developed or wider sections.

Those at the bottom (e.g., Channel 36, Channel 46, Channel 72) represent steeper stretches or sections with different dynamics, with lower velocities than the gradient would suggest, often linked to high roughness or the presence of structures.

Component loadings of the variables on the factor plane



This graph is the factorial plot of variables (biplot) on components 1 and 2: each arrow represents a variable projected onto the space defined by these two axes.

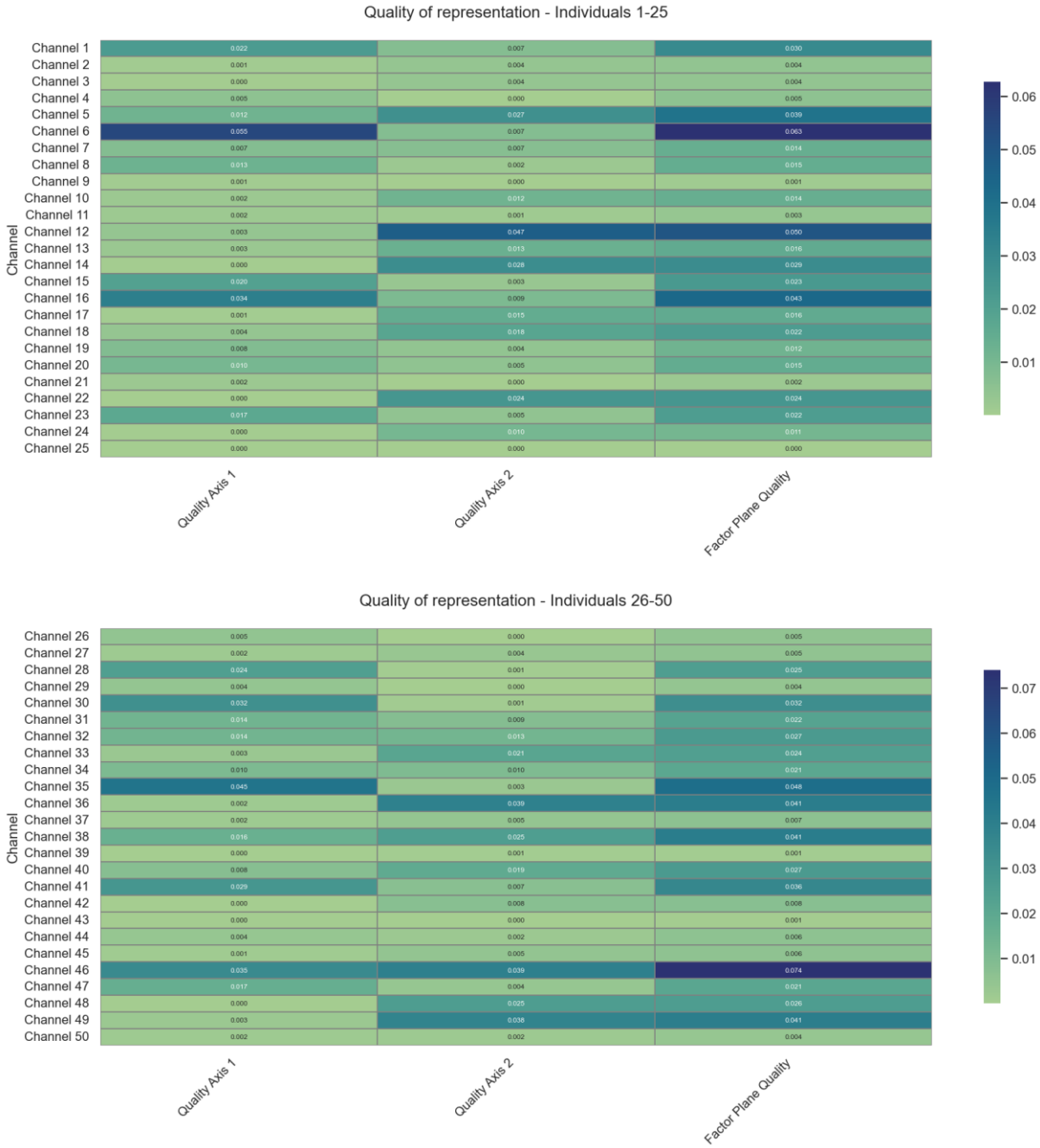
Axis Direction

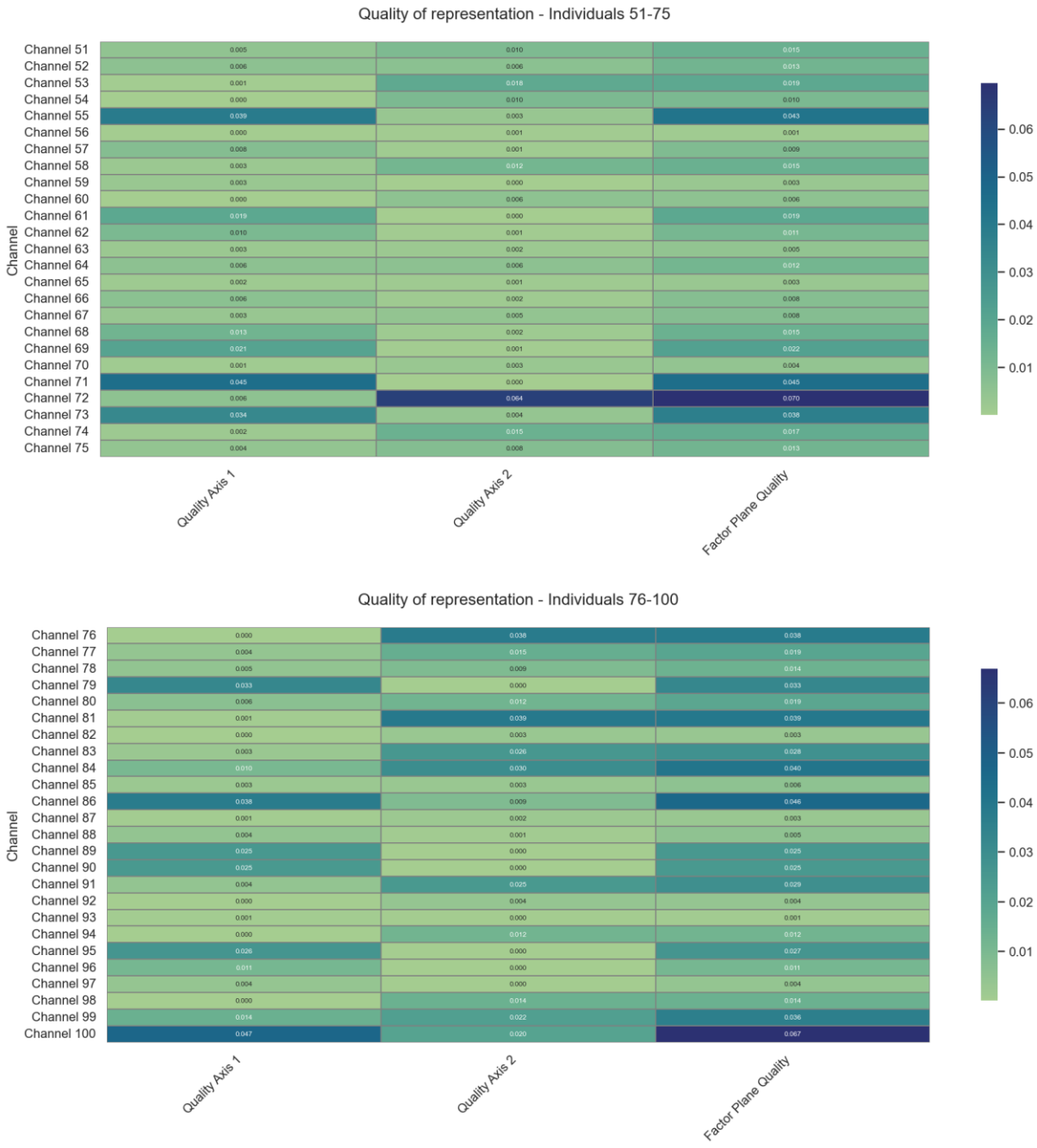
- On axis 1 (horizontal), the arrows for Depth, Roughness, Siltation, and to a lesser extent Temperature point to the right, while Flow Rate and Width point to the left.
    - Right side: deep, rough, warm, and silted channels.
    - Left side: smoother, shallower, and less silted channels.
- On axis 2 (vertical), the Velocity arrow points upwards, while Slope and to a lesser extent Flow Rate/Width point downwards.
    - At the top: high velocity for a relatively moderate slope.
    - At the bottom: steeper slope but lower velocity.

Relationships between variables

- Arrows pointing in similar directions (for example, Roughness and Siltation) indicate positively correlated variables: they often increase or decrease together.
- Nearly opposite arrows (Speed vs Slope, or Flow/Width vs Roughness/Siltation) suggest negative correlations in this respect: when one increases, the other tends to decrease.

Quality of representation relative to axis 1, axis 2, and the factor plane





This heatmap shows the quality of channel representation on axes 1 and 2, and on the factorial plane (1–2), i.e., the cos^2 values of the PCA scores for each individual.

Meaning of the values

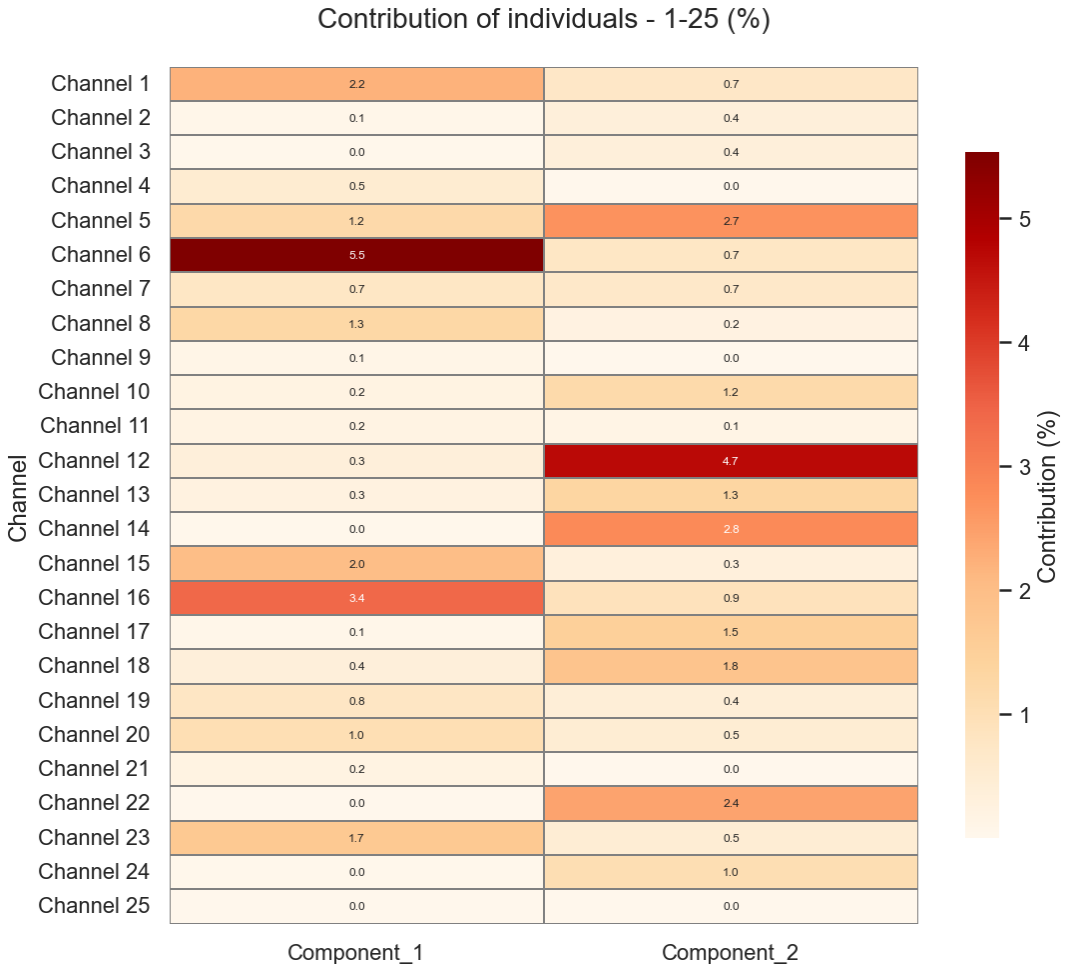
- Each row corresponds to a channel, each column to the quality on axis 1, the quality on axis 2, and then the quality on the plane (axes 1–2).
- The values range from 0 to 1: the higher the value (darker color), the more reliable and well-represented the channel's position on the axis or plane.

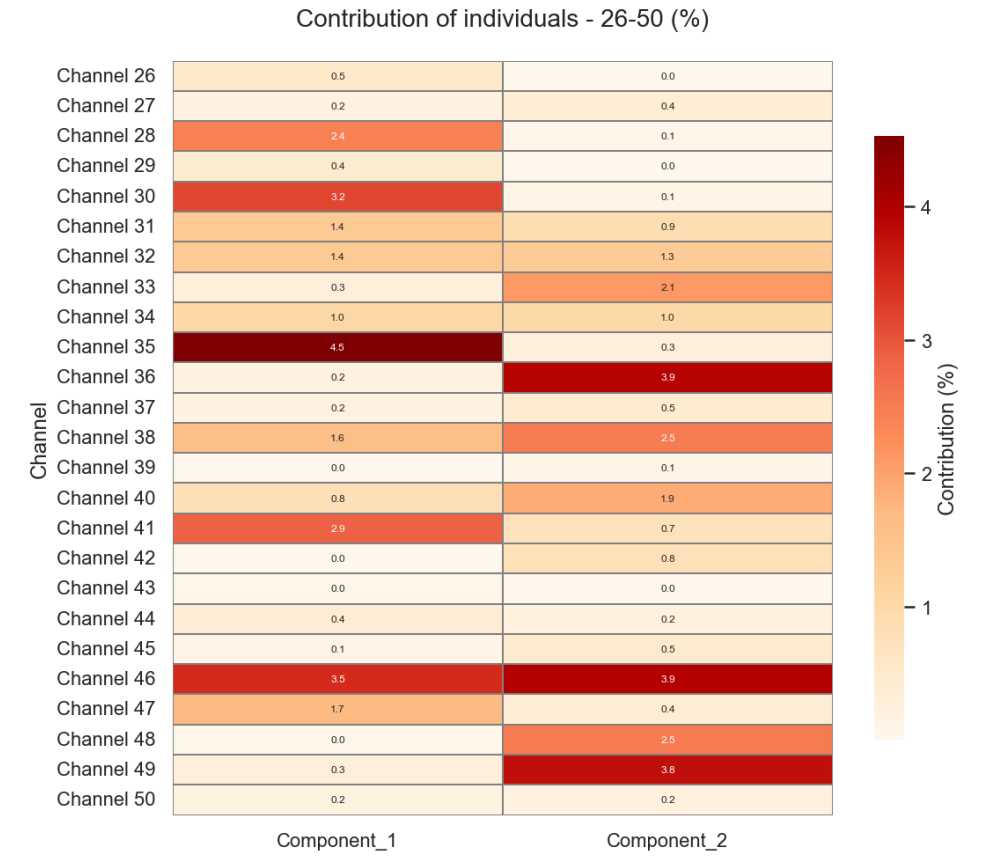
Practical interpretation

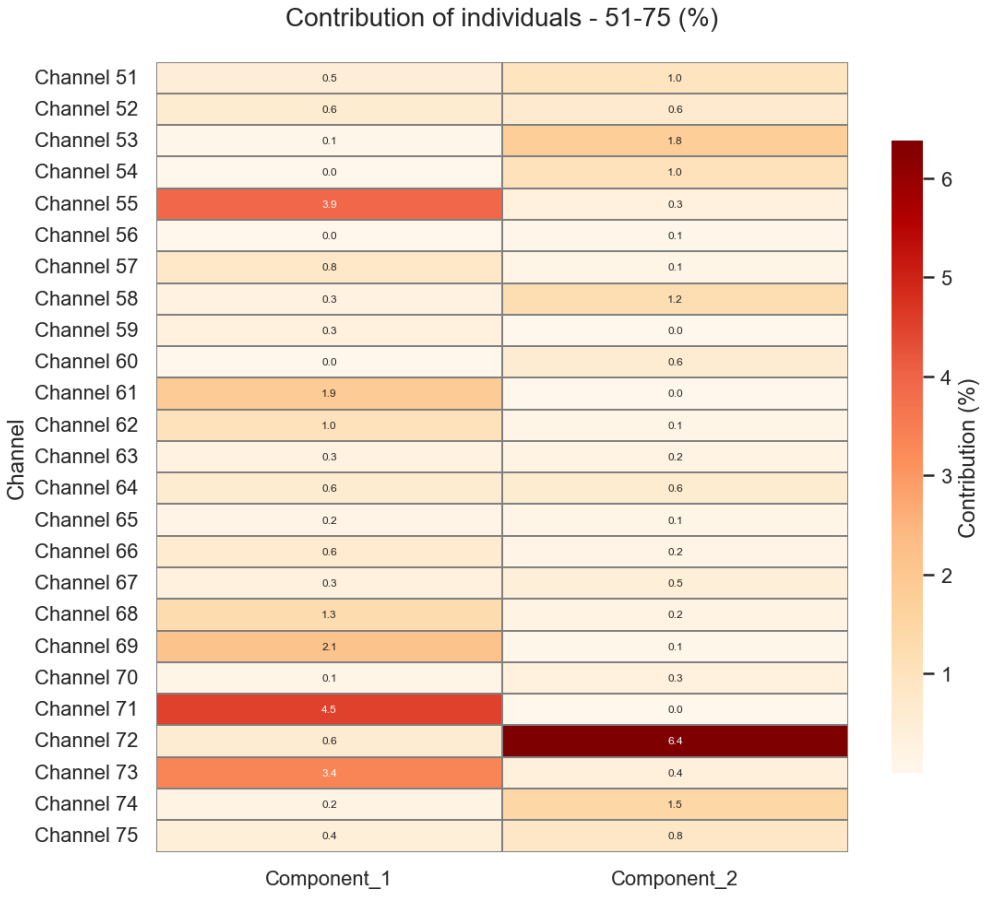
- Most cells are in light green shades with low values (a few hundredths): this confirms that axes 1 and 2 explain only a modest portion of the total variance, as you saw in the inertia table (≈38%).
- Some channels (for example, around numbers 46, 72, 100) show slightly stronger values in a column or on the plane, indicating that they are better aligned with axes 1–2 and therefore more "typical" of the contrasts summarized by these axes.
- Conversely, channels with very low values everywhere are not well described by the 1–2 plane: for them, the information lies more in components 25, 56, etc., and their position on the factorial plane should be interpreted with caution.

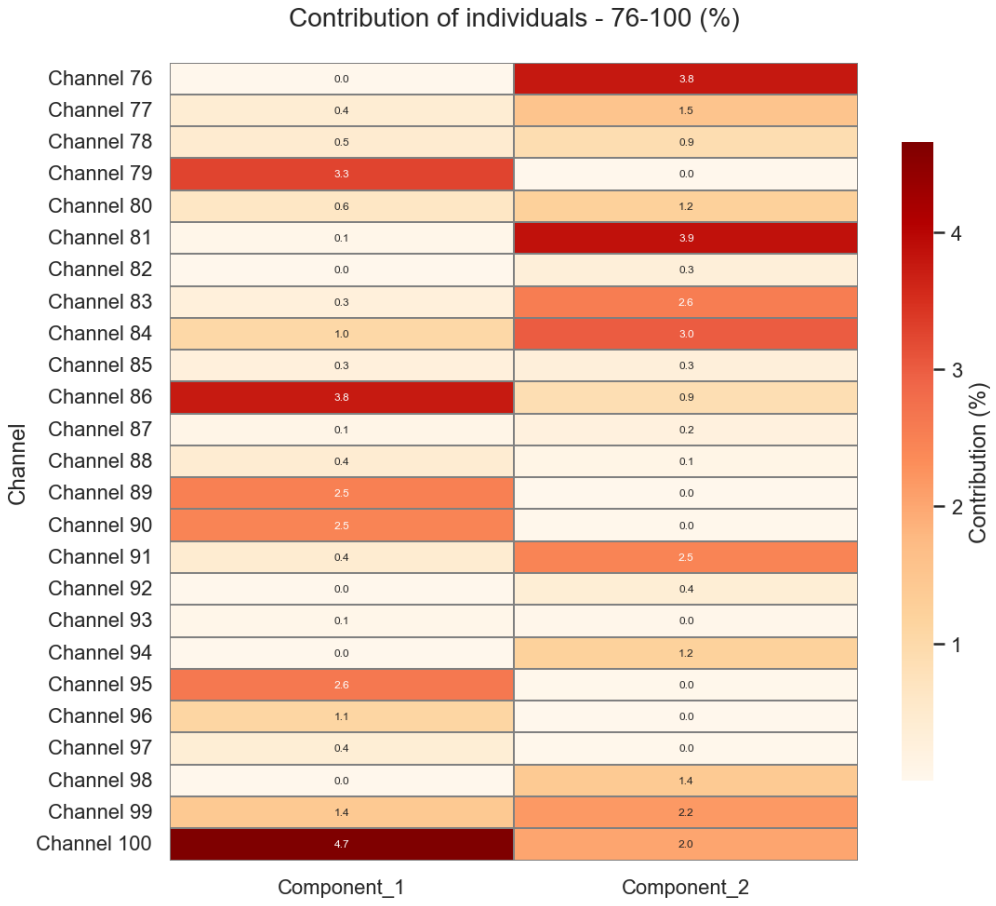
In summary, this figure helps to judge the degree to which we can trust the projection of each channel onto the (C1, C2) plane: the higher the $\cos^2$, the more relevant the graphical interpretation of that channel on this plane.

Contribution of individuals and variables




Contribution of individuals - 26-50 (%)
Channel 26 0.5 0.0
Channel 27 0.2 0.4
Channel 28 2.4 0.1
Channel 29 0.4 0.0
Channel 30 3.2 0.1
Channel 31 1.4 0.9
Channel 32 1.4 1.3
Channel 33 0.3 2.1
Channel 34 1.0 1.0
Channel 35 4.5 0.3
Channel 36 0.2 3.9
Channel 37 0.2 0.5
Channel 38 1.6 2.5
Channel 39 0.0 0.1
Channel 40 0.8 1.9
Channel 41 2.9 0.7
Channel 42 0.0 0.8
Channel 43 0.0 0.0
Channel 44 0.4 0.2
Channel 45 0.1 0.5
Channel 46 3.5 3.9
Channel 47 1.7 0.4
Channel 48 0.0 2.5
Channel 49 0.3 3.8
Channel 50 0.2 0.2
Channel
Component_1
Component_2
Contribution (%)
1
2
3
4


Contribution of individuals - 51-75 (%)
Channel 51 0.5 1.0
Channel 52 0.6 0.6
Channel 53 0.1 1.8
Channel 54 0.0 1.0
Channel 55 3.9 0.3
Channel 56 0.0 0.1
Channel 57 0.8 0.1
Channel 58 0.3 1.2
Channel 59 0.3 0.0
Channel 60 0.0 0.6
Channel 61 1.9 0.0
Channel 62 1.0 0.1
Channel 63 0.3 0.2
Channel 64 0.6 0.6
Channel 65 0.2 0.1
Channel 66 0.6 0.2
Channel 67 0.3 0.5
Channel 68 1.3 0.2
Channel 69 2.1 0.1
Channel 70 0.1 0.3
Channel 71 4.5 0.0
Channel 72 0.6 6.4
Channel 73 3.4 0.4
Channel 74 0.2 1.5
Channel 75 0.4 0.8
Channel
Component_1
Component_2
Contribution (%)
1
2
3
4
5
6



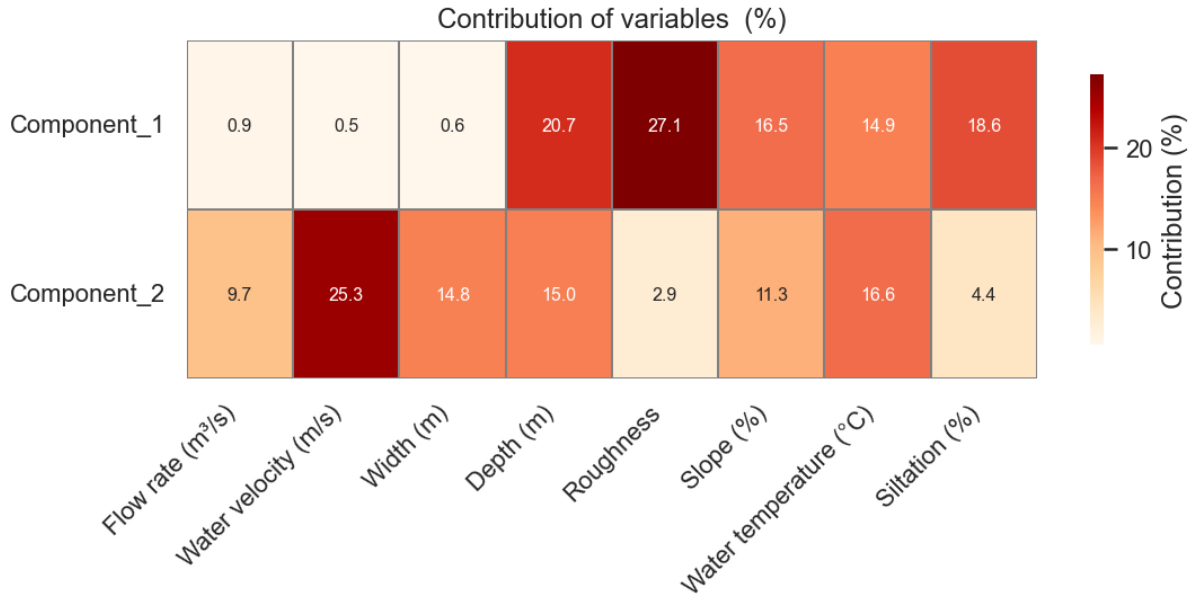
This graph shows the contribution of each channel to the construction of axes 1 and 2 of the PCA (as a percentage).

What the contribution means

- Each row = one channel, each column = its contribution to Component 1 or Component 2.
- The darker the cell and the higher the value (e.g., > 3–4%), the more weight that channel carries in defining the axis: these are the "driving" individuals of the axis, those that pull the scatter plot in each direction.

Interpreting the key areas

- For Component 1, a few channels clearly stand out (e.g., Channel 6, Channel 35, Channel 71, Channel 100, etc.) with contributions significantly higher than the theoretical average (100/100 = 1% if everyone contributed equally).
    - These channels are those that best embody the "roughness-depth-siltation" gradient represented by axis 1: they are located at the far right or far left of the factorial plane. •
- For Component 2, other channels dominate (e.g., Channel 46, Channel 72, and some neighboring channels) with stronger contributions.
    - They are responsible for the "high velocity/lower slope" versus "steep slope/lower velocity" contrast that structures axis 2.



This table shows the contribution of each variable to the construction of axes 1 and 2 of the PCA, as a percentage of the axis variance.

Component 1

- The variables that strongly structure axis 1 are, in order, roughness (27.1%), depth (20.7%), siltation (18.6%), slope (16.5%), and temperature (14.9%).
- Flow rate, velocity, and width contribute very little (≤1%), which confirms that Component 1 is essentially a "channel condition and sedimentation" axis: the rougher, deeper, warmer, and more silted a channel is, the higher its score on C1.

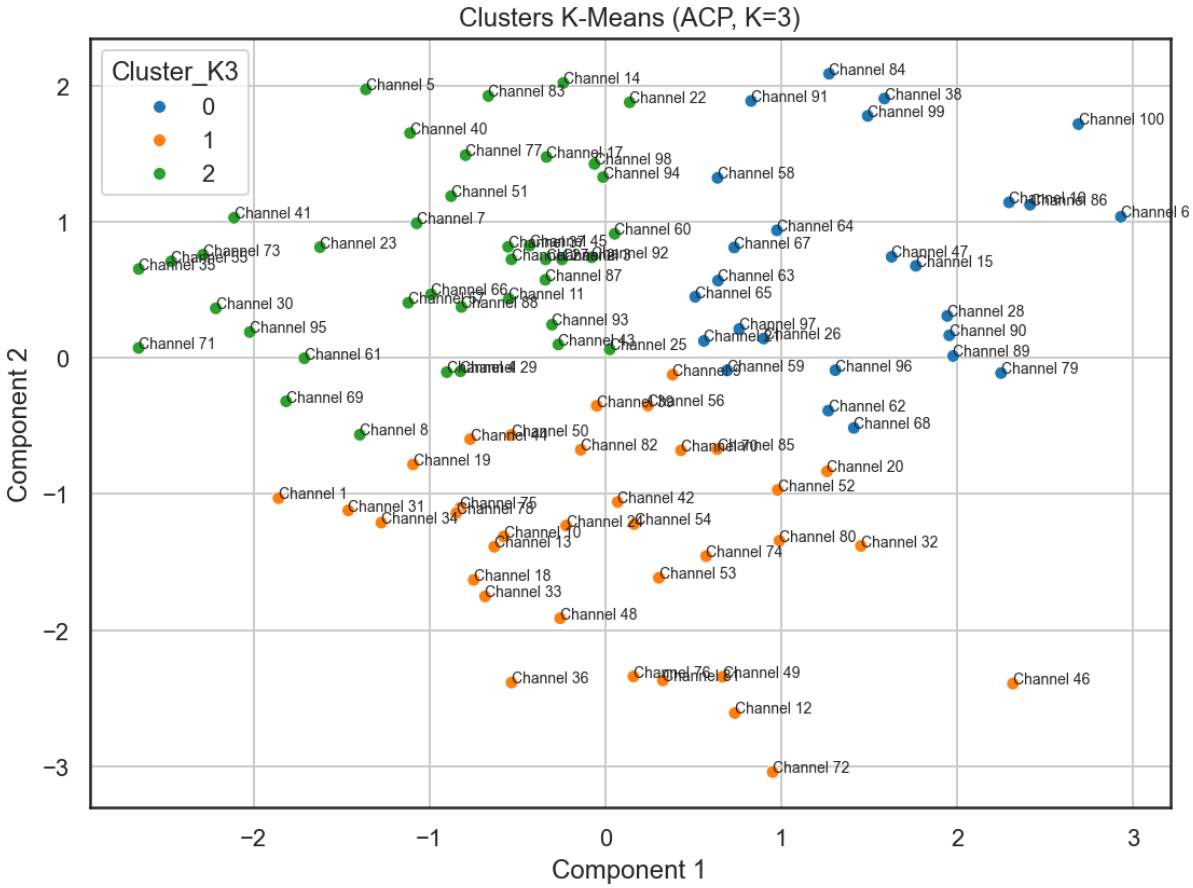
Component 2

- Axis 2 is dominated by water velocity (25.3%), followed by temperature (16.6%), width (14.8%), depth (15.0%), slope (11.3%), and flow rate (9.7%).
- Roughness and siltation contribute little (≈3% and 4.4%, respectively), indicating that C2 describes a flow dynamic gradient (velocity, width, depth, slope) rather than a sedimentation gradient.

Chapter II: Clustering with KMeans and Random Forest Modeling

Part 1: Clustering with KMeans

K-Means application with K = 3



General Meaning

- Each point represents a channel, positioned by its scores on C1 (roughness–depth–siltation) and C2 (velocity–slope), and colored according to its cluster (0, 1, or 2).
- The clusters group together hydraulically similar channels within this limited area: channels in the same group have similar profiles in terms of channel condition, sedimentation, velocity, and slope.

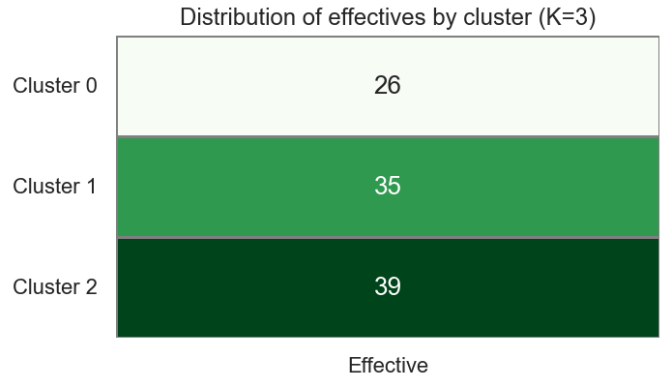
Reading the Three Clusters (Typical Interpretation)

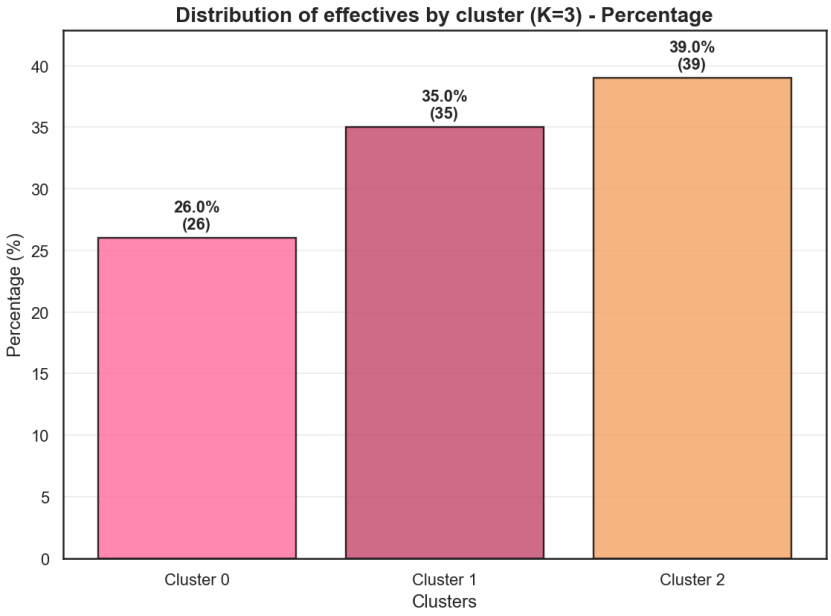
Based on the interpretation of the axes:

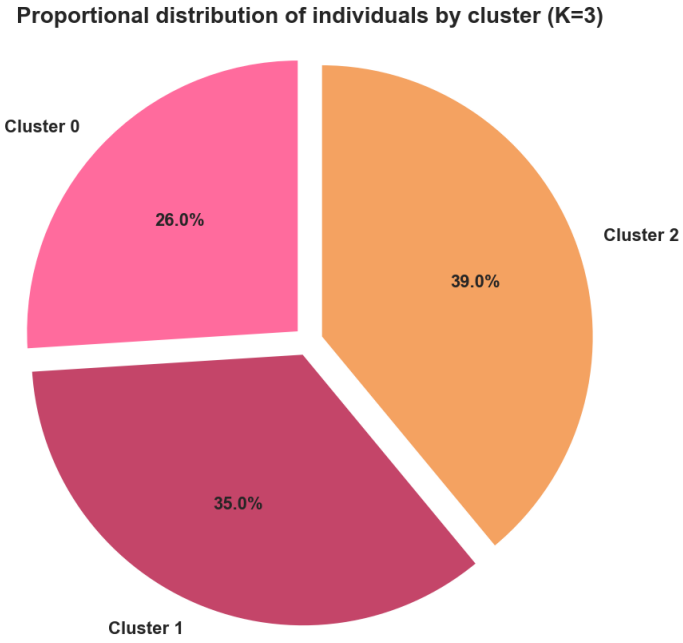
- Cluster 0 (blue): primarily on the right side of the map, often in a medium to high position on C2. This group includes rough, deep, fairly silted channels with moderate to high velocities and rather low to medium slopes. These are sediment-laden sections, morphologically distinct, but with still significant flow dynamics.

- Cluster 1 (orange): Located towards the bottom of the graph, with negative C2 scores, and distributed along axis 1. These channels have lower velocities for often steeper gradients, reflecting sections with reduced flow (roughness, structures, head losses) or storage sections. They represent a hydraulically “slower” profile, sometimes with steep theoretical gradients but limited flow.

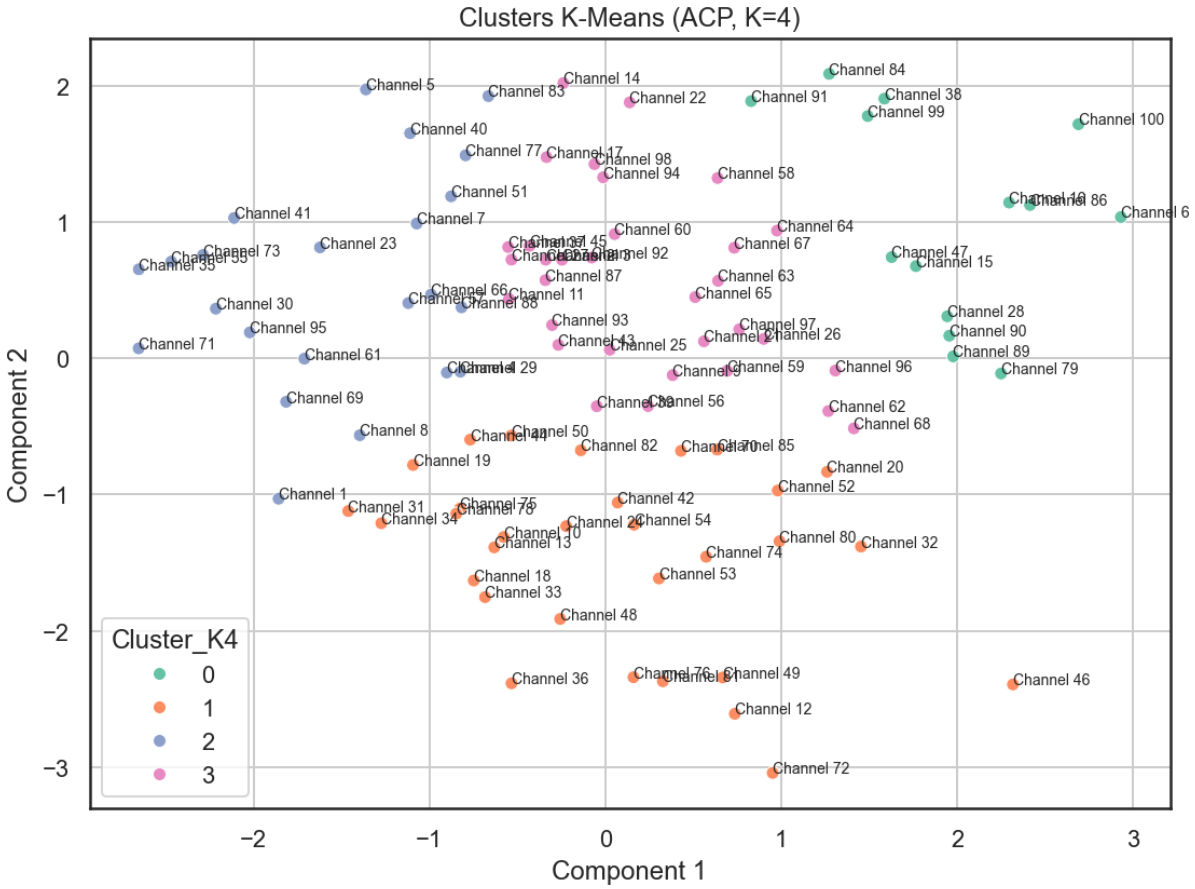
- Cluster 2 (green): Located towards the left and center/top of the map. These are less silted and less rough channels, often with a “cleaner” or maintained profile, moderate velocities, and variable gradients. They may correspond to more artificial/stabilized channels, where the channel condition is less degraded.




Distribution of effectives by cluster (K=3) - Percentage
Percentage (%)
0
5
10
15
20
25
30
35
40
26.0%
(26)
35.0%
(35)
39.0%
(39)
Cluster 0
Cluster 1
Cluster 2
Clusters


Proportional distribution of individuals by cluster (K=3)
Cluster 0
26.0%
Cluster 2
39.0%
35.0%
Cluster 1

K-Means application with K = 4



This graph presents the K-means clustering with K = 4 applied to the channels, projected onto the factorial plane (components 1 and 2).

General logic

- Each point represents a channel positioned according to its scores on C1 (channel condition/siltation, to the right) and C2 (speed-slope dynamics, upwards).
- The colors represent the four groups of channels with similar hydraulic profiles within this limited area.

Interpretation of the 4 clusters

Following the axis directions:

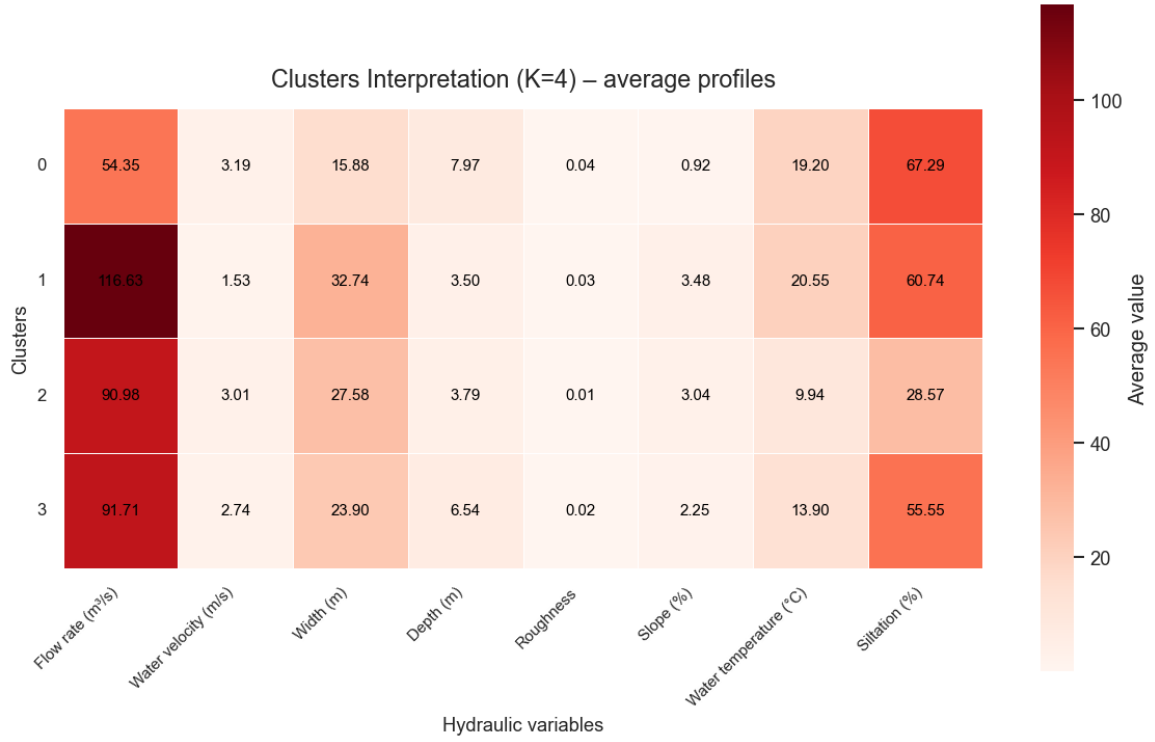
- Cluster 0 (green)
    - Located mainly on the right and rather high or in the center.
    - Rough, deep, fairly silted channels with moderate to high velocities and rather low to medium slopes.
    - Profile: morphologically loaded but hydraulically active sections.
- Cluster 1 (orange)
    - Mostly at the bottom of the graph, sometimes very low (Channels 36, 46, 72).

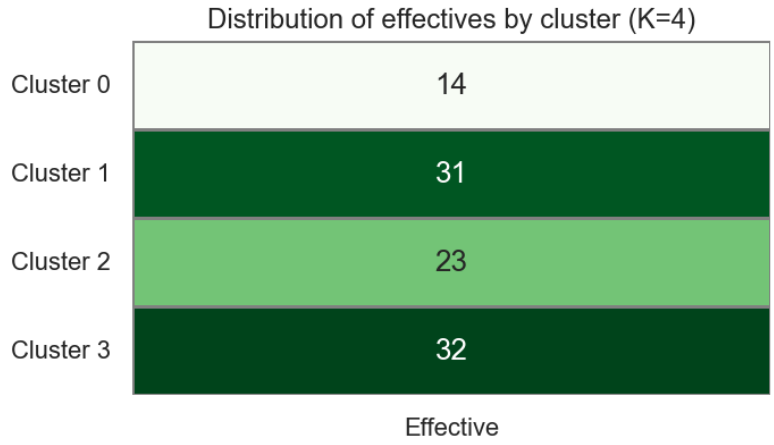
- Channels with negative C2 scores: steep gradients but relatively low velocities, possibly due to high roughness or structures.
- Profile: slowed/storage sections, despite significant gradients.

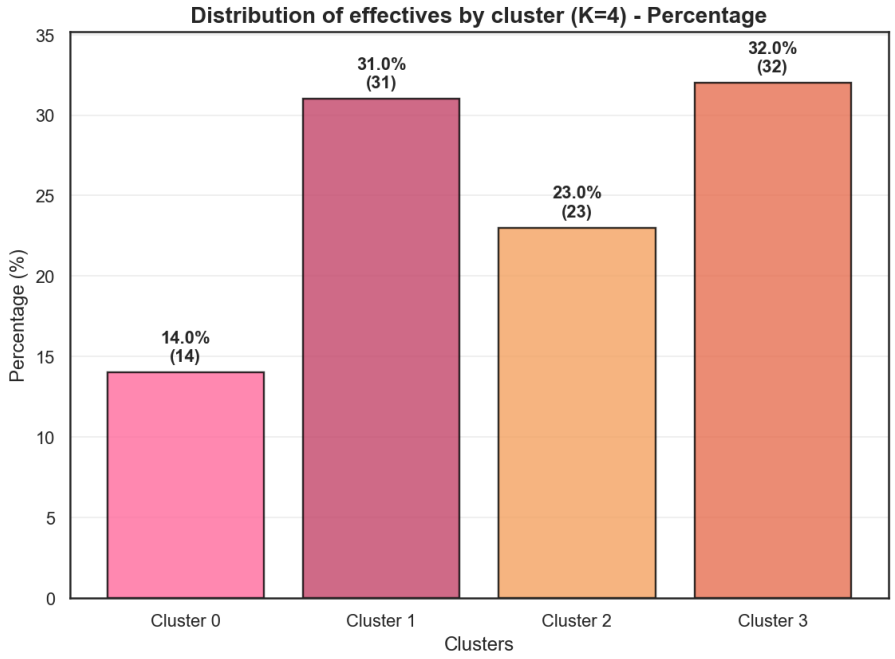
- Cluster 2 (blue)
  - Mostly on the left and in the center/upper part.
  - Channels with less siltation, less roughness, closer to an "average" state on C1, with moderate velocities.
  - Profile: "cleaner" or maintained channels, with less extreme morphology.
- Cluster 3 (pink/purple)
  - Grouped mainly around the center right, with a slightly positive C2 score.
  - Moderately silted and rough channels, but with slightly higher velocities, intermediate between clusters 0 and 2.
  - Profile: intermediate sections, of the transitional type between heavily silted and more artificially modified channels.

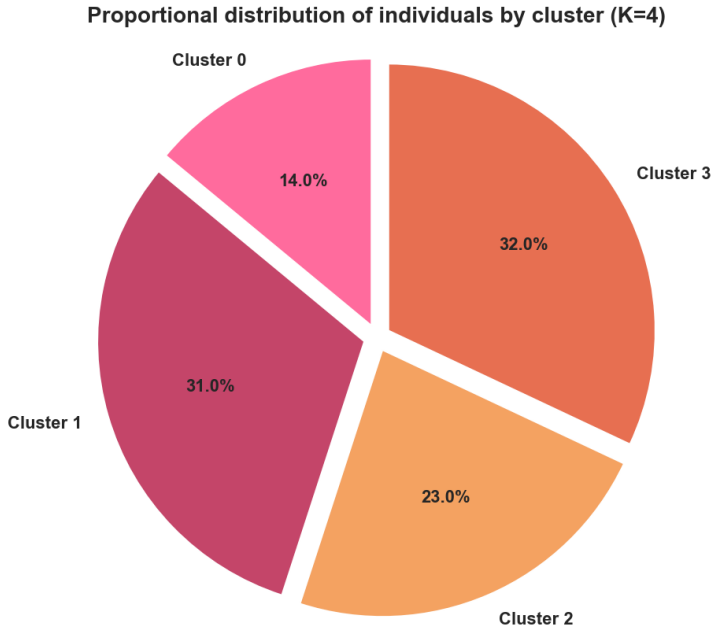
The advantage of switching from K=3 to K=4

With K = 4, our dataset is more finely divided: the “mud-bound/complex” group of the K = 3 case is now subdivided into subtypes (for example cluster 0 vs 3), which allows a more detailed typology of the channels.

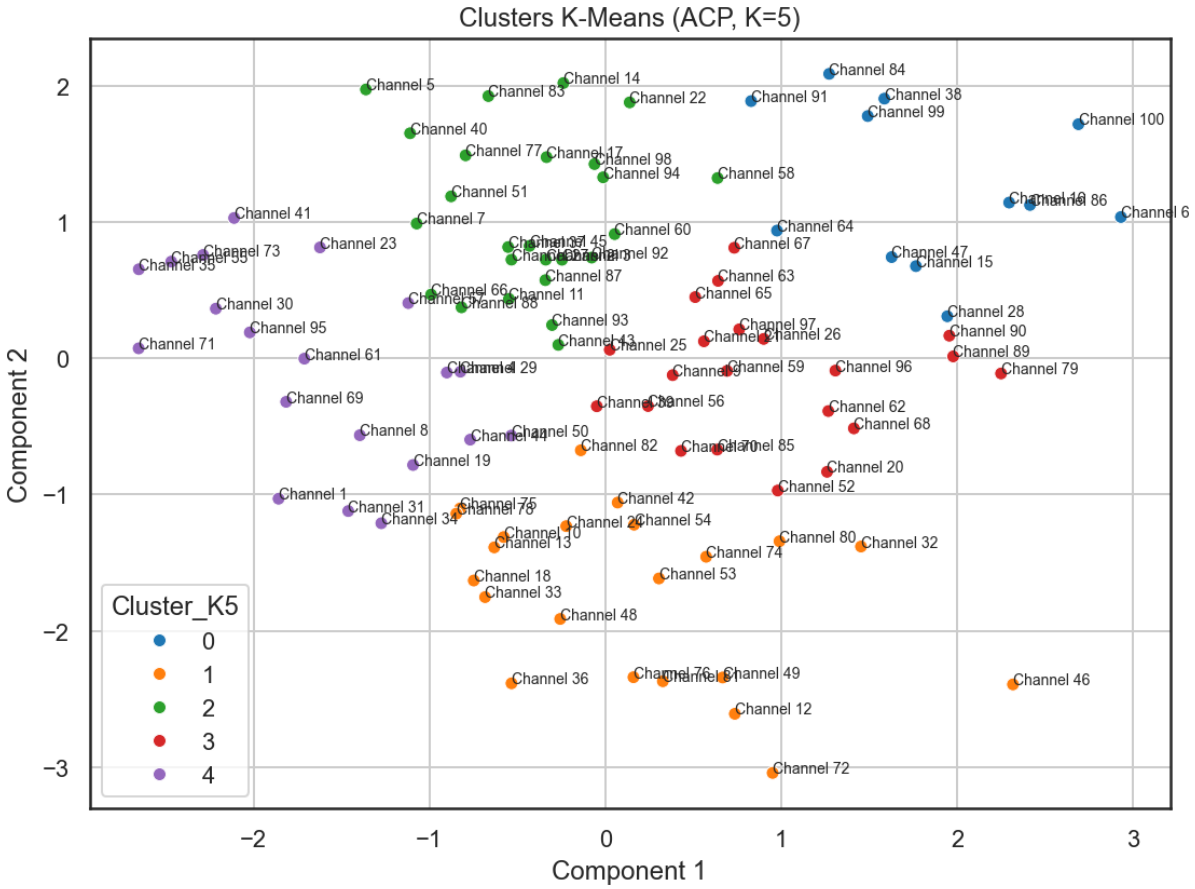



Distribution of effectives by cluster (K=4)
Cluster 0
Cluster 1
Cluster 2
Cluster 3
14
31
23
32
Effective


Distribution of effectives by cluster (K=4) - Percentage
Percentage (%)
0
5
10
15
20
25
30
35
14.0%
(14)
31.0%
(31)
23.0%
(23)
32.0%
(32)
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Clusters


Proportional distribution of individuals by cluster (K=4)
Cluster 0
14.0%
Cluster 3
32.0%
31.0%
Cluster 1
23.0%
Cluster 2

K-Means application with K = 5



This graph shows the K-means clusters with K = 5 on the PCA factorial plane (axes 1 and 2).

As before:

- Axis 1 = "bed state/siltation" gradient (right = rough, deep, silted).
- Axis 2 = "flow dynamics" gradient (top = high velocity for moderate slope; bottom = steep slope but lower velocity).

Interpretation of the 5 groups (typical profile)

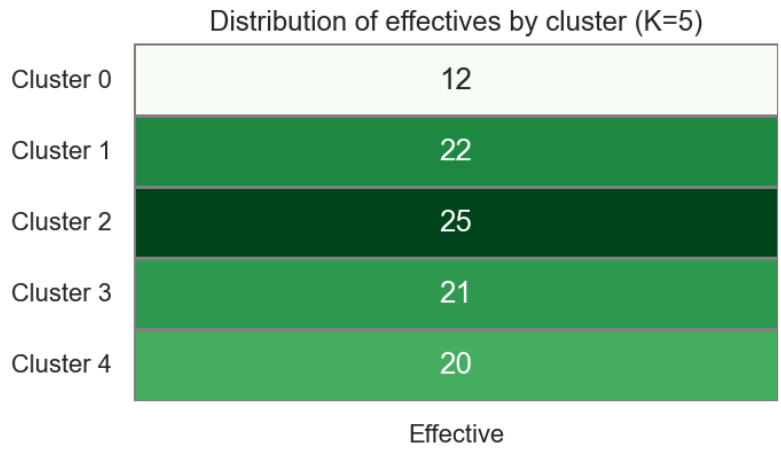
To keep it concise, they can be interpreted as follows:

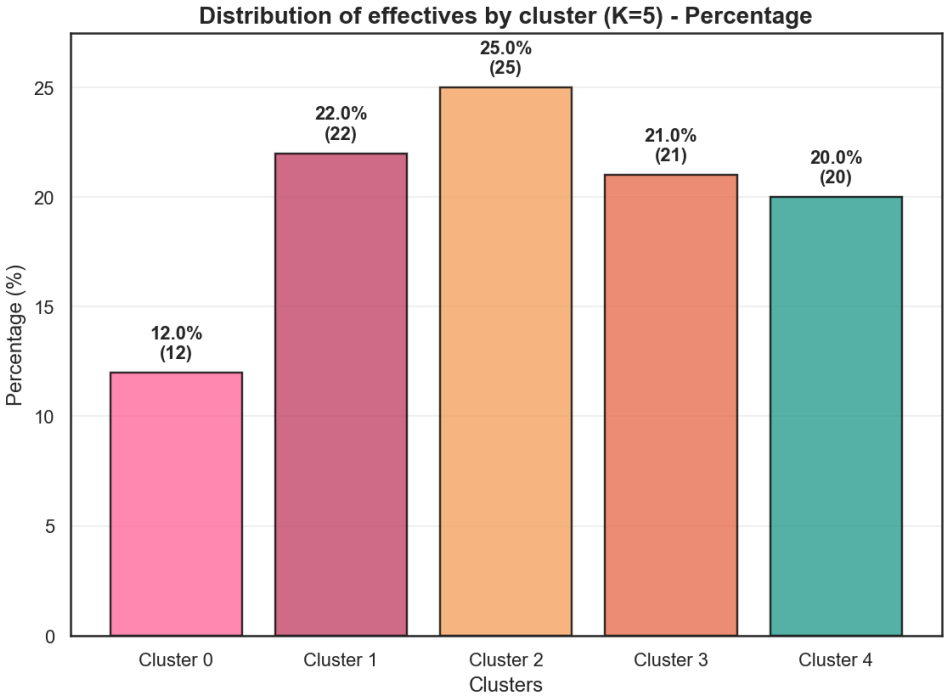
- Cluster 0 (blue)
    - On the right and quite high.
    - Very silted/rough, deep channels with moderate to high velocities.
    - Sediment-laden sections that are still dynamic, often with a moderate slope.
- Cluster 1 (orange)
    - Mostly at the bottom (negative C2), sometimes very eccentric (Channels 36, 46, 72).
    - Channels with steep gradients or restricted conditions, lower than expected velocities, are often more extreme in depth or geometry.

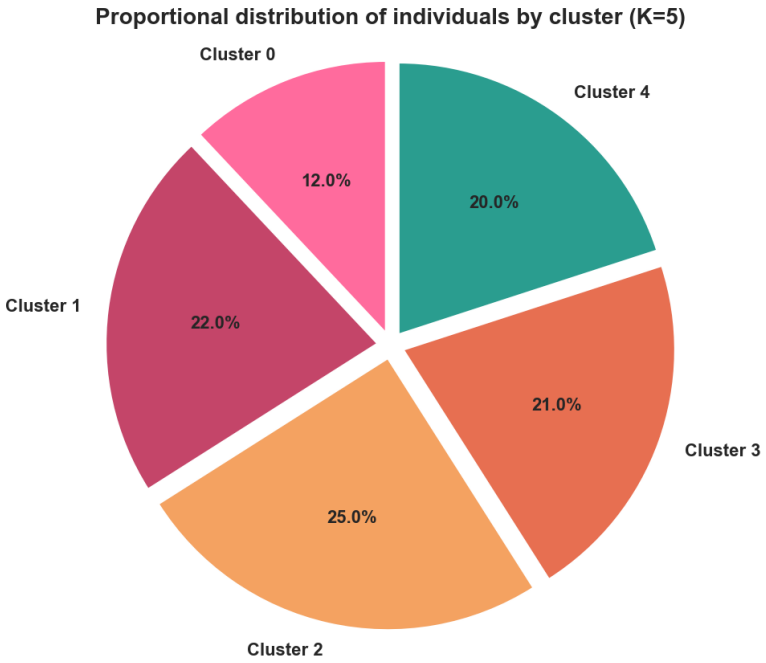
    - Profile: torrential or heavily restricted/constructed sections, requiring monitoring for stability.
- Cluster 2 (green)
    - Located primarily in the center-right and at the top.
    - Deep channels, fairly silted, but with more regular flow dynamics (acceptable velocities, moderate gradients).
    - Profile: morphologically distinct but "functional" sections, intermediate between clusters 0 and 3/4.
- Cluster 3 (red)
    - In the center-right but lower than cluster 2.
    - Silted/rough channels but with lower velocities, sometimes in storage areas or with calm flow regimes.
    - Profile: silted sections with slow dynamics, where the risk of continuous sediment accumulation is significant.
- Cluster 4 (purple)
    - Primarily to the left, often near the center or slightly at the top.
    - Channels with less siltation and less roughness, close to average for speed and gradient.
    - Profile: “cleaner” or improved channels, relatively stable morphology.

Importance of K = 5

The shift to 5 clusters allows for a more precise distinction between different forms of siltation and dynamics (for example, separating heavily silted but fast-flowing channels from heavily silted and slow-flowing ones).







Part 2: Modeling with Radom Forest

→ Flow rate (m^3/s):

- MSE train: 177.08, test: 1307.54
- R2 train: 0.943, test: 0.708

→ Water velocity (m/s):

- MSE train: 0.11, test: 0.92
- R2 train: 0.948, test: 0.580

→ Width (m):

- MSE train: 8.37, test: 91.51

· R2 train: 0.957, test: 0.590
→ Depth (m):
· MSE train: 0.36, test: 3.86
· R2 train: 0.949, test: 0.547
→ Roughness:
· MSE train: 0.00, test: 0.00
· R2 train: 0.916, test: 0.596
→ Slope (%):
· MSE train: 0.14, test: 1.00
· R2 train: 0.939, test: 0.511
→ Water temperature (°C):
· MSE train: 4.25, test: 30.14
· R2 train: 0.896, test: 0.422
→ Siltation (%):
· MSE train: 67.73, test: 790.90
· R2 train: 0.912, test: 0.028

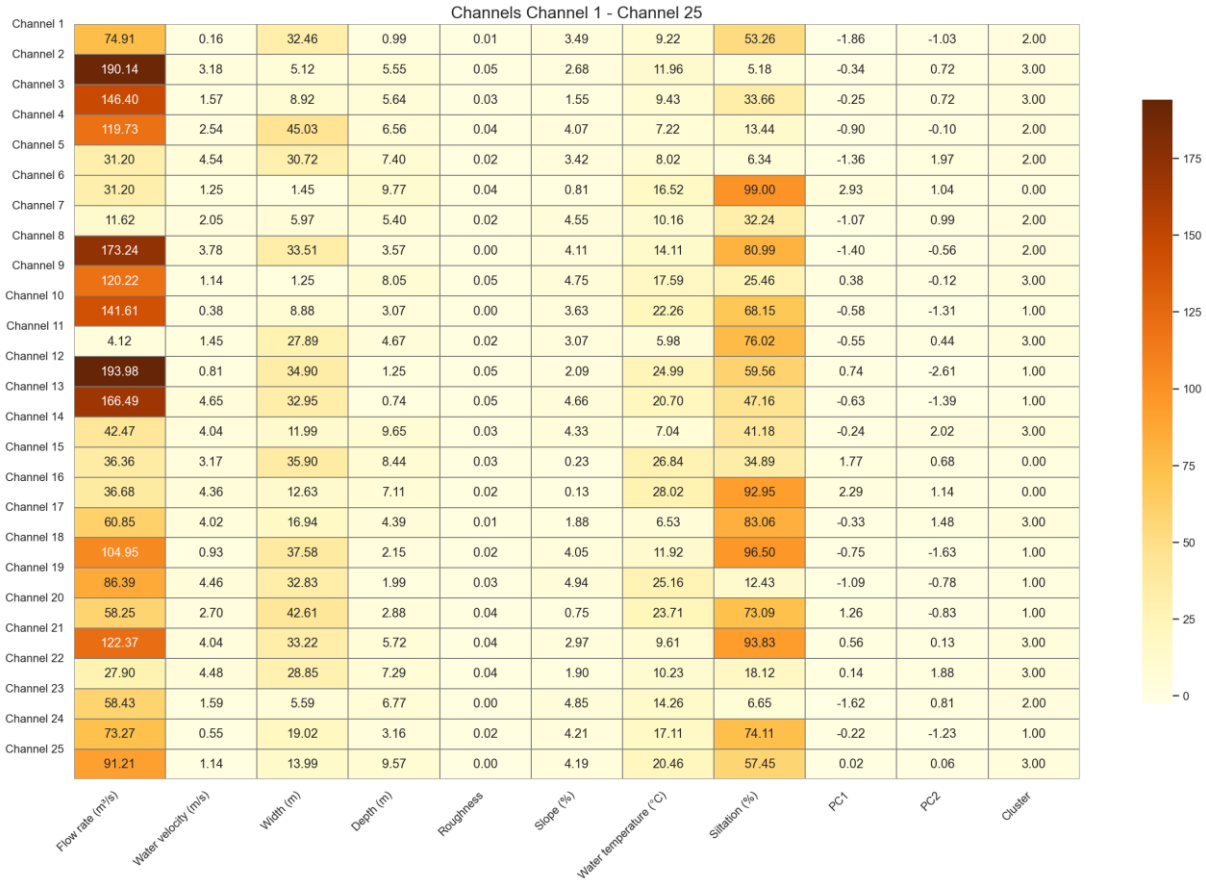
Overall Model Quality

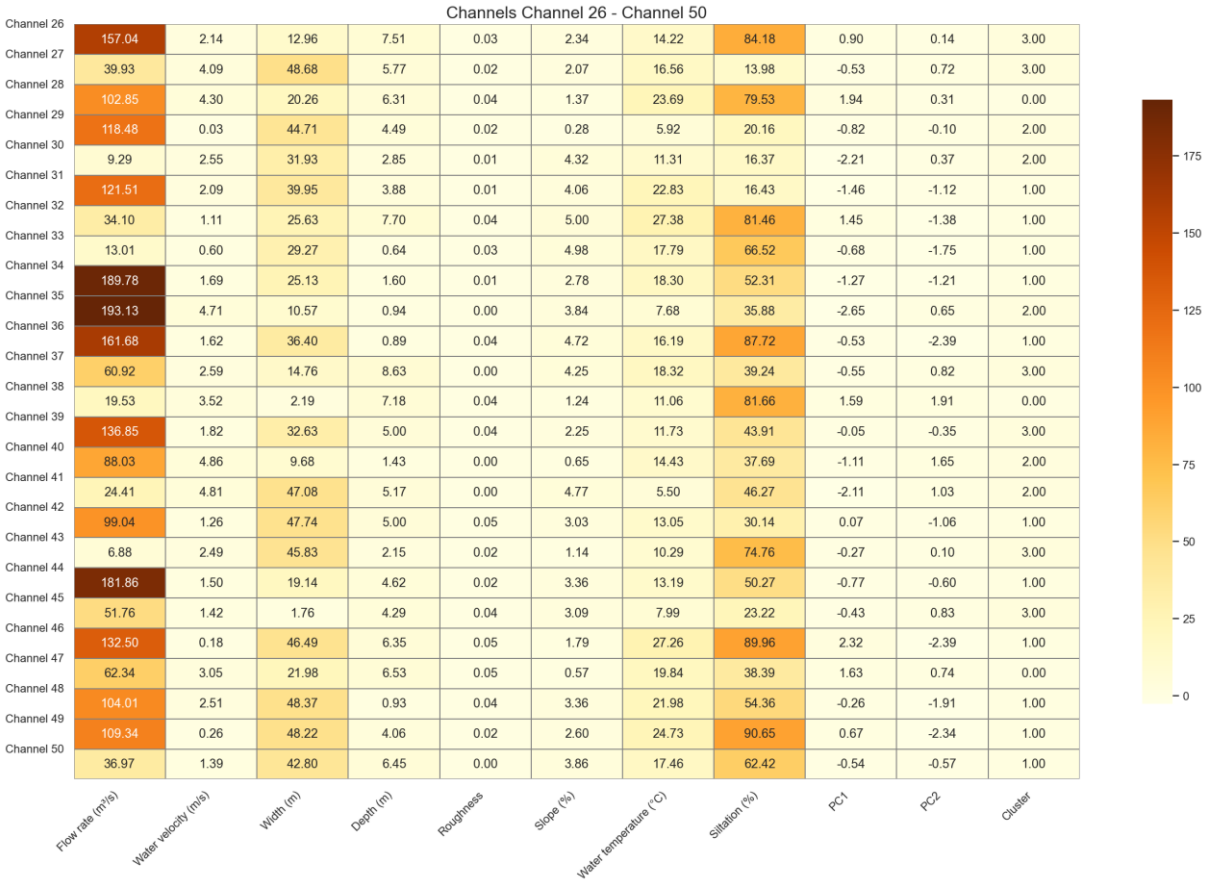
- For most variables, the R^2 training value is high (≈0.90–0.96): the model explains the variance well in the training data.
- In the test model, the R^2 values remain average to acceptable (≈0.5–0.7) for flow rate, velocity, width, depth, roughness, slope, and temperature: the model retains predictive power, but with a significant loss compared to the training model, reflecting some overfitting and the complexity of the phenomenon.

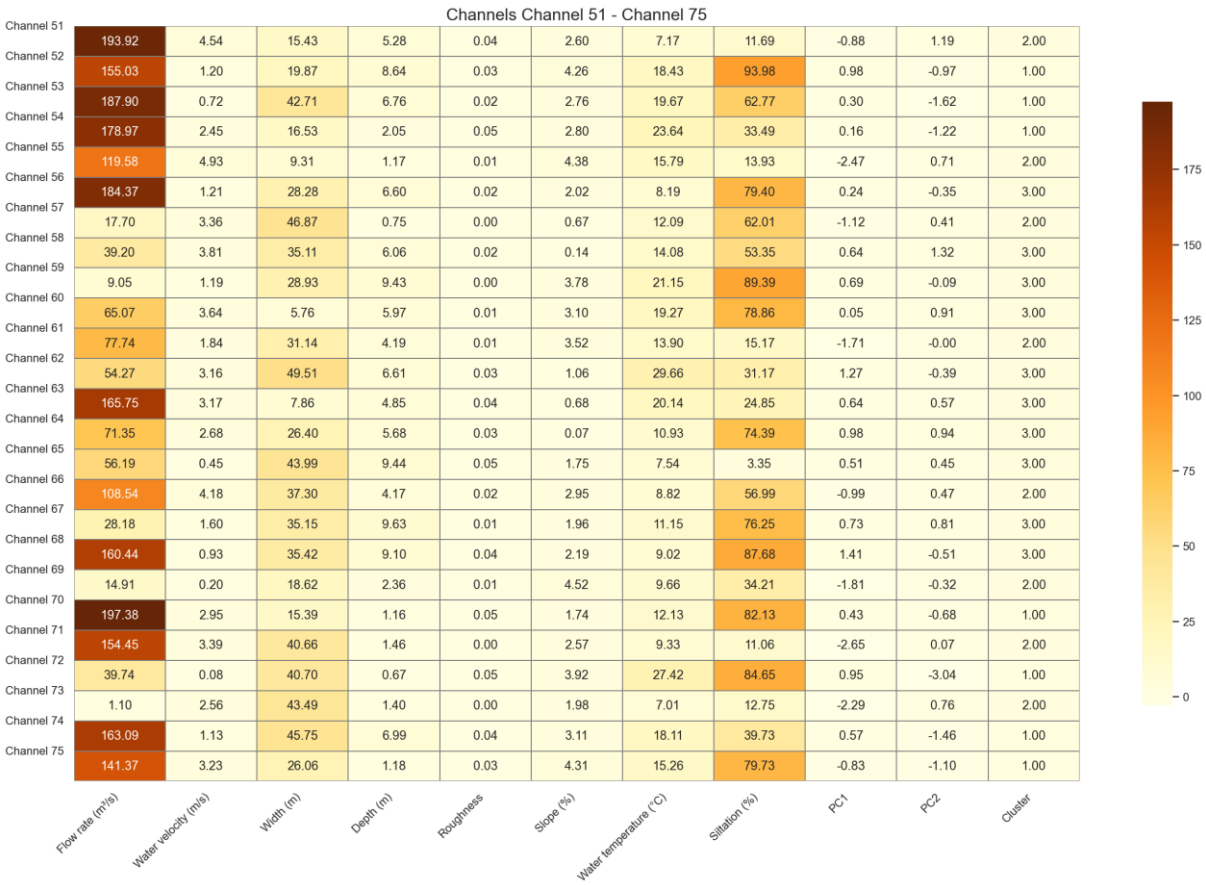
Variable by variable

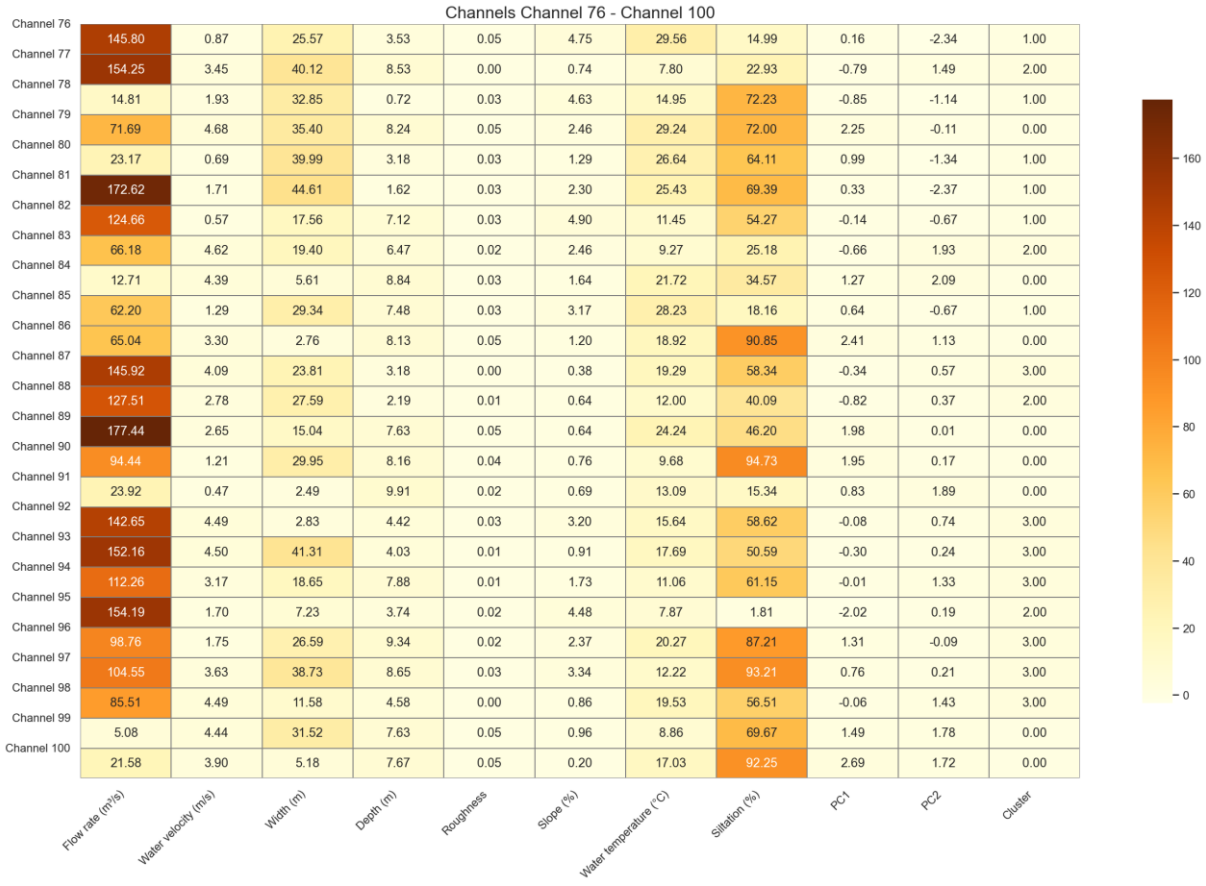
- Flow rate (m^3/s): R^2 training = 0.943, test = 0.708. The model reconstructs the flow rate fairly well, but the MSE error is almost 7 times greater between the training and test models (177 → 1308), which shows that extreme flow rates or certain channels are poorly predicted in generalization.
- Water velocity (m/s): R^2 training = 0.948, test = 0.580. The model captures large velocity trends well during training, but loses accuracy on new channels, probably because velocity is highly dependent on nonlinear combinations of slope, roughness, and geometry.
- Width (m): R^2 training = 0.957, test = 0.590. Very good fit during training, but significant degradation during testing; width is more difficult to predict at unseen sites.
- Depth (m): R^2 training = 0.949, test = 0.547. Similar behavior to width: good modeling of depths in the training sample, but limited extrapolation capacity.
- Roughness: R^2 training = 0.916, test = 0.596, with very low MSE. Roughness is predicted fairly well, suggesting that it is closely linked to other characteristics (geometry, siltation, slope), even though performance decreases in testing.

- Slope (%): R^2 training = 0.939, test = 0.511. The model generally identifies the slope categories, but the error remains significant on the new channels, likely because the slope varies considerably locally.
- Water temperature (°C): R^2 training = 0.896, test = 0.422. Temperature is less well explained: it depends on external factors (season, climate, shading) that are not all included in the input variables, hence the average predictive power.
- Siltation (%): R^2 training = 0.912, test = 0.028. This is the most problematic variable: good explanation in training, but virtually no variance explained in testing. The model does not generalize siltation; it mainly memorizes the values of the train, which is consistent with a very local phenomenon and difficult to predict (solid transport, deposits, management of structures).




Channels Channel 26 - Channel 50
Channel 26 157.04 2.14 12.96 7.51 0.03 2.34 14.22 84.18 0.90 0.14 3.00
Channel 27 39.93 4.09 48.68 5.77 0.02 2.07 16.56 13.98 -0.53 0.72 3.00
Channel 28 102.85 4.30 20.26 6.31 0.04 1.37 23.69 79.53 1.94 0.31 0.00
Channel 29 118.48 0.03 44.71 4.49 0.02 0.28 5.92 20.16 -0.82 -0.10 2.00
Channel 30 9.29 2.55 31.93 2.85 0.01 4.32 11.31 16.37 -2.21 0.37 2.00
Channel 31 121.51 2.09 39.95 3.88 0.01 4.06 22.83 16.43 -1.46 -1.12 1.00
Channel 32 34.10 1.11 25.63 7.70 0.04 5.00 27.38 81.46 1.45 -1.38 1.00
Channel 33 13.01 0.60 29.27 0.64 0.03 4.98 17.79 66.52 -0.68 -1.75 1.00
Channel 34 189.78 1.69 25.13 1.60 0.01 2.78 18.30 52.31 -1.27 -1.21 1.00
Channel 35 193.13 4.71 10.57 0.94 0.00 3.84 7.68 35.88 -2.65 0.65 2.00
Channel 36 161.68 1.62 36.40 0.89 0.04 4.72 16.19 87.72 -0.53 -2.39 1.00
Channel 37 60.92 2.59 14.76 8.63 0.00 4.25 18.32 39.24 -0.55 0.82 3.00
Channel 38 19.53 3.52 2.19 7.18 0.04 1.24 11.06 81.66 1.59 1.91 0.00
Channel 39 136.85 1.82 32.63 5.00 0.04 2.25 11.73 43.91 -0.05 -0.35 3.00
Channel 40 88.03 4.86 9.68 1.43 0.00 0.65 14.43 37.69 -1.11 1.65 2.00
Channel 41 24.41 4.81 47.08 5.17 0.00 4.77 5.50 46.27 -2.11 1.03 2.00
Channel 42 99.04 1.26 47.74 5.00 0.05 3.03 13.05 30.14 0.07 -1.06 1.00
Channel 43 6.88 2.49 45.83 2.15 0.02 1.14 10.29 74.76 -0.27 0.10 3.00
Channel 44 181.86 1.50 19.14 4.62 0.02 3.36 13.19 50.27 -0.77 -0.60 1.00
Channel 45 51.76 1.42 1.76 4.29 0.04 3.09 7.99 23.22 -0.43 0.83 3.00
Channel 46 132.50 0.18 46.49 6.35 0.05 1.79 27.26 89.96 2.32 -2.39 1.00
Channel 47 62.34 3.05 21.98 6.53 0.05 0.57 19.84 38.39 1.63 0.74 0.00
Channel 48 104.01 2.51 48.37 0.93 0.04 3.36 21.98 54.36 -0.26 -1.91 1.00
Channel 49 109.34 0.26 48.22 4.06 0.02 2.60 24.73 90.65 0.67 -2.34 1.00
Channel 50 36.97 1.39 42.80 6.45 0.00 3.86 17.46 62.42 -0.54 -0.57 1.00
Flow rate (m³/s)
Water velocity (m/s)
Width (m)
Depth (m)
Roughness
Slope (%)
Water temperature (°C)
Siltation (%)
PC1
PC2
Cluster
175
150
125
100
75
50
25
0


Channels Channel 51 - Channel 75
Channel 51 193.92 4.54 15.43 5.28 0.04 2.60 7.17 11.69 -0.88 1.19 2.00
Channel 52 155.03 1.20 19.87 8.64 0.03 4.26 18.43 93.98 0.98 -0.97 1.00
Channel 53 187.90 0.72 42.71 6.76 0.02 2.76 19.67 62.77 0.30 -1.62 1.00
Channel 54 178.97 2.45 16.53 2.05 0.05 2.80 23.64 33.49 0.16 -1.22 1.00
Channel 55 119.58 4.93 9.31 1.17 0.01 4.38 15.79 13.93 -2.47 0.71 2.00
Channel 56 184.37 1.21 28.28 6.60 0.02 2.02 8.19 79.40 0.24 -0.35 3.00
Channel 57 17.70 3.36 46.87 0.75 0.00 0.67 12.09 62.01 -1.12 0.41 2.00
Channel 58 39.20 3.81 35.11 6.06 0.02 0.14 14.08 53.35 0.64 1.32 3.00
Channel 59 9.05 1.19 28.93 9.43 0.00 3.78 21.15 89.39 0.69 -0.09 3.00
Channel 60 65.07 3.64 5.76 5.97 0.01 3.10 19.27 78.86 0.05 0.91 3.00
Channel 61 77.74 1.84 31.14 4.19 0.01 3.52 13.90 15.17 -1.71 -0.00 2.00
Channel 62 54.27 3.16 49.51 6.61 0.03 1.06 29.66 31.17 1.27 -0.39 3.00
Channel 63 165.75 3.17 7.86 4.85 0.04 0.68 20.14 24.85 0.64 0.57 3.00
Channel 64 71.35 2.68 26.40 5.68 0.03 0.07 10.93 74.39 0.98 0.94 3.00
Channel 65 56.19 0.45 43.99 9.44 0.05 1.75 7.54 3.35 0.51 0.45 3.00
Channel 66 108.54 4.18 37.30 4.17 0.02 2.95 8.82 56.99 -0.99 0.47 2.00
Channel 67 28.18 1.60 35.15 9.63 0.01 1.96 11.15 76.25 0.73 0.81 3.00
Channel 68 160.44 0.93 35.42 9.10 0.04 2.19 9.02 87.68 1.41 -0.51 3.00
Channel 69 14.91 0.20 18.62 2.36 0.01 4.52 9.66 34.21 -1.81 -0.32 2.00
Channel 70 197.38 2.95 15.39 1.16 0.05 1.74 12.13 82.13 0.43 -0.68 1.00
Channel 71 154.45 3.39 40.66 1.46 0.00 2.57 9.33 11.06 -2.65 0.07 2.00
Channel 72 39.74 0.08 40.70 0.67 0.05 3.92 27.42 84.65 0.95 -3.04 1.00
Channel 73 1.10 2.56 43.49 1.40 0.00 1.98 7.01 12.75 -2.29 0.76 2.00
Channel 74 163.09 1.13 45.75 6.99 0.04 3.11 18.11 39.73 0.57 -1.46 1.00
Channel 75 141.37 3.23 26.06 1.18 0.03 4.31 15.26 79.73 -0.83 -1.10 1.00
Flow rate (m³/s)
Water velocity (m/s)
Width (m)
Depth (m)
Roughness
Slope (%)
Water temperature (°C)
Siltation (%)
PC1
PC2
Cluster
175
150
125
100
75
50
25
0



As part of enriching our analysis, we have integrated a new set of channels characterized by quantitative variables such as Flow Rate (m³/s), Water Velocity (m/s), Width (m), Depth (m), Roughness, Slope (%), and Water Temperature (°C). This data is summarized in a clear table for ease of understanding.

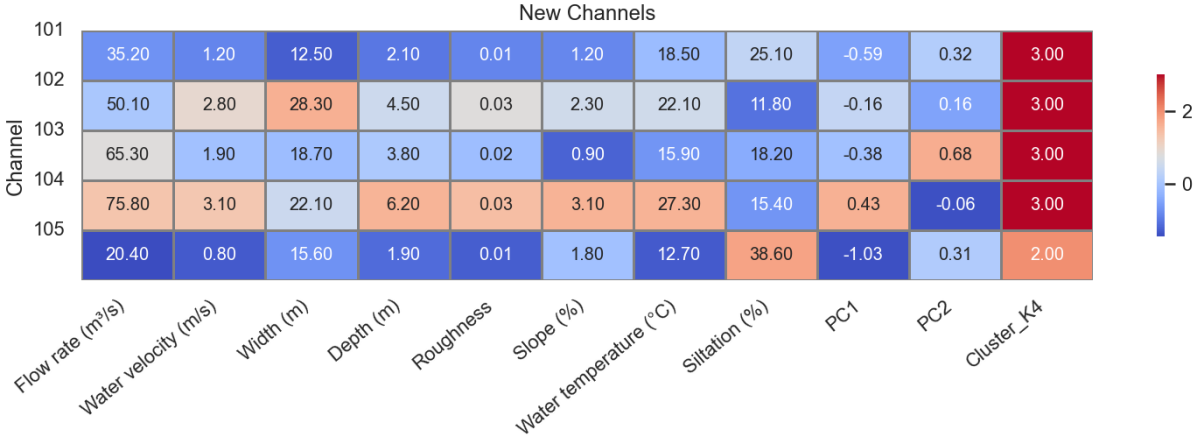
The objective of this step is to apply our initially defined statistical processing chain: data transformation by centering and reduction (using the previously trained scaler), projection of principal component space via a pre-established PCA model, and then assignment of a cluster to each new channel based on the K-Means clustering model trained on the historical data.

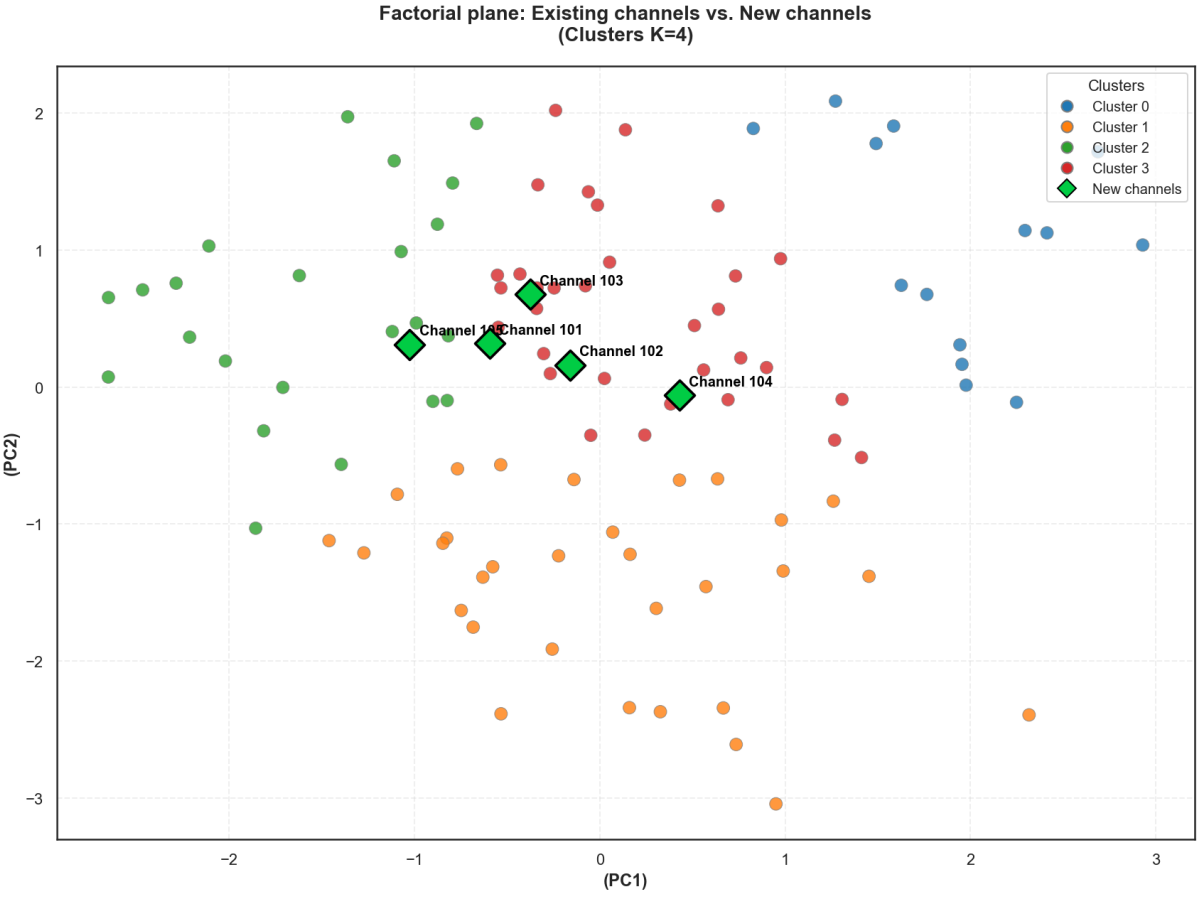
This process allows us to objectively assess the position of the new channels relative to existing clusters, anticipate their typology, and provide key elements for decision-making and future data use, in accordance with asset lifecycle methodologies and best practices in statistical analysis.

This positioning is visualized in the factorial plane derived from PCA, where the old and new channels are represented simultaneously, with a clear distinction between the assigned clusters. This representation facilitates the interpretation of the results and the integration of new projects into our overall monitoring.

Channel	Flow rate (m³/s)	Water velocity (m/s)	Width (m)	Depth (m)	Roughness	Slope (%)	Water temperature (°C)	Siltation (%)
101	35.2	1.2	12.5	2.1	0.012	1.2	18.5	25.1

102	50.1	2.8	28.3	4.5	0.028	2.3	22.1	11.8
103	65.3	1.9	18.7	3.8	0.019	0.9	15.9	18.2
104	75.8	3.1	22.1	6.2	0.034	3.1	27.3	15.4
105	20.4	0.8	15.6	1.9	0.008	1.8	12.7	38.6





This graph shows the PCA factorial plane (PC1–PC2) of existing channels, colored by cluster (K = 4), and the position of 4 new channel projects (101 to 105) represented by green diamonds.

Reading the axes and clusters

- The horizontal axis (PC1) primarily summarizes the gradient of channel condition/siltation/roughness: to the right, channels with more silt and roughness; to the left, "cleaner" channels.
- The vertical axis (PC2) summarizes the flow dynamics (velocity–slope): at the top, channels with higher velocity and moderate slope; at the bottom, steeper slopes but lower velocities.
- The colors indicate the 4 families of existing channels (Cluster 0, 1, 2, 3) that we have already interpreted in our previous analysis (e.g., heavily silted channels, slow and steep channels, more maintained channels, etc.).

Position of the new channels

- The new channels 101–104 are located in the center right of the plan, in an area already occupied by channels from the red cluster (Cluster 3) and partly close to the red channels except 105 closes to the green channels.
- Their position indicates that they exhibit:
    - A moderate to fairly high level of siltation/roughness (PC1 > 0),
    - Average flow dynamics (PC2 around 0–0.7), neither overtly torrential nor particularly slow.

Interpretation for the project

- Channels 101–105 do not create a new hydraulic type; they are morphologically and hydraulically related to channel classes already observed in the network (mainly cluster 3).

We can therefore conclude that, from the point of view of the overall characteristics (geometry, roughness, siltation, flow/speed), these new projects are consistent with the existing ones and comparable to canals already in operation, which facilitates the anticipation of their behavior (risk of siltation, flow regime, maintenance needs).

Chapter III: Analysis of Weighting Data and (CA)

Contingency Table

Channel	Flow rate (m^3/s)	Water velocity (m/s)	Width (m)	Depth (m)	Roughness	Slope (%)	Water temperature (°C)	Siltation (%)
Channel 1	7	4	8	5	7	10	3	7
Channel 2	8	5	4	8	8	3	6	5
Channel 3	2	8	6	2	5	1	10	6
Channel 4	9	1	10	3	7	4	9	3
Channel 5	5	3	7	5	9	7	2	4
Channel 6	9	2	10	9	10	5	2	4

Channel 7	7	8	3	1	4	2	8	4
Channel 8	2	6	6	10	4	6	2	10

Frequency Matrix P

This figure represents a matrix of normalized frequencies for 8 channels (1 to 8) and 8 criteria (flow rate, velocity, width, depth, roughness, slope, temperature, siltation).

What the values mean

- Each cell contains a small value (≈0.003 to 0.028): this is the relative frequency or weight of the criterion for a given channel, after row/column normalization.
- The higher the value and the darker the color, the more dominant this criterion is for the channel in question compared to the other criteria in the matrix.

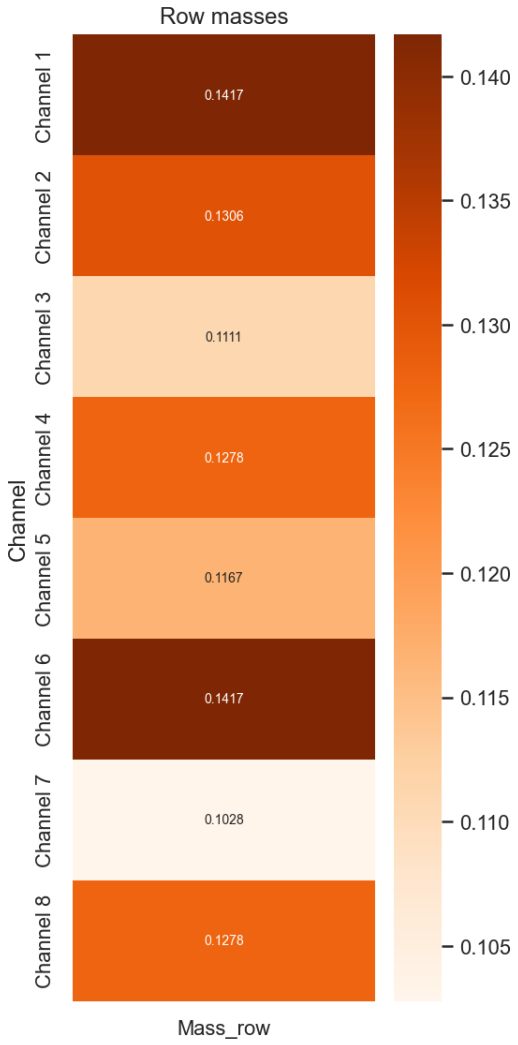
Interpretation by examples

- For Channel 4, the highest values are for width (0.028), flow rate (0.025), and temperature (0.025): these criteria contribute more to its characterization.
- For Channel 6, the highest frequencies are found in flow rate (0.025), width (0.028), depth (0.025), and roughness (0.028): this channel is characterized by a large gauge and a very prominent bed.
- For Channel 3, velocity (0.022) and temperature (0.028) stand out more, suggesting a channel characterized more by its flow dynamics and thermal context than by its dimensions.
- For Channel 8, depth (0.028) and siltation (0.028) dominate, suggesting a deep and heavily silted channel, despite more moderate values for the other criteria.

Usefulness for our analysis

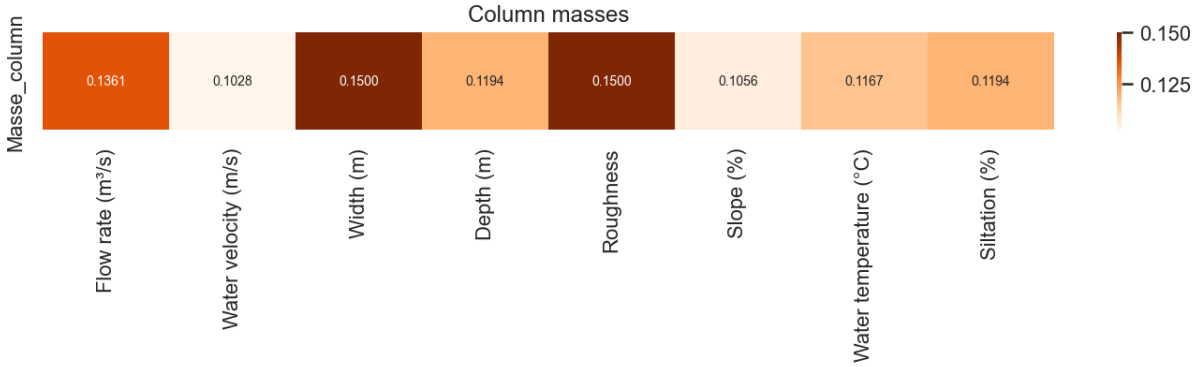
This matrix allows us to quickly compare which criteria are the most discriminating from one channel to another and to identify profiles: channels dominated by geometry (width/depth), by dynamics (speed/slope) or by the state of the bed (roughness/siltation).

Row and column masses



- Channel 1 – Mass 0,1417
  Channel 1 has the highest mass, meaning it contributes significantly to the overall structure of the frequency matrix. Its alone accounts for nearly 14% of the total "importance" of the 8 channels and therefore plays a central role in the analysis.
- Channel 2 – Mass 0,1306
  Channel 2 also occupies a significant position, with a mass slightly lower than that of Channel 1. It significantly influences the results and can be considered one of the most representative channels in the dataset.
- Channel 3 – Mass 0,1111
  The mass of Channel 3 is more moderate, indicating a decent but less dominant contribution. It participates in the overall structure without being a pivotal channel.

- Channel 4 – Mass 0,1278
  Channel 4 has a relatively high mass, close to that of Channel 2. It plays a significant role in defining the average profiles and factorial axes resulting from the analysis.
- Channel 5 – Mass 0,1167
  With an intermediate mass, channel 5 contributes in a balanced way to the whole. It is neither marginal nor dominant but clearly participates in the data structure.
- Channel 6 – Mass 0,1417
  Like channel 1, channel 6 has the highest mass. It is therefore a particularly influential channel, whose profile weighs heavily in the overall results and in the extracted axes.
- Channel 7 – Mass 0,1028
  Channel 7 has the lowest mass in the sample. Its contribution to the overall structure is limited, suggesting a more marginal or less representative channel, though not negligible.
- Channel 8 – Mass 0,1278
  Channel 8 has a mass comparable to that of channel 4. It participates significantly in the analysis and can be considered a medium to high-weight channel in the overall balance.

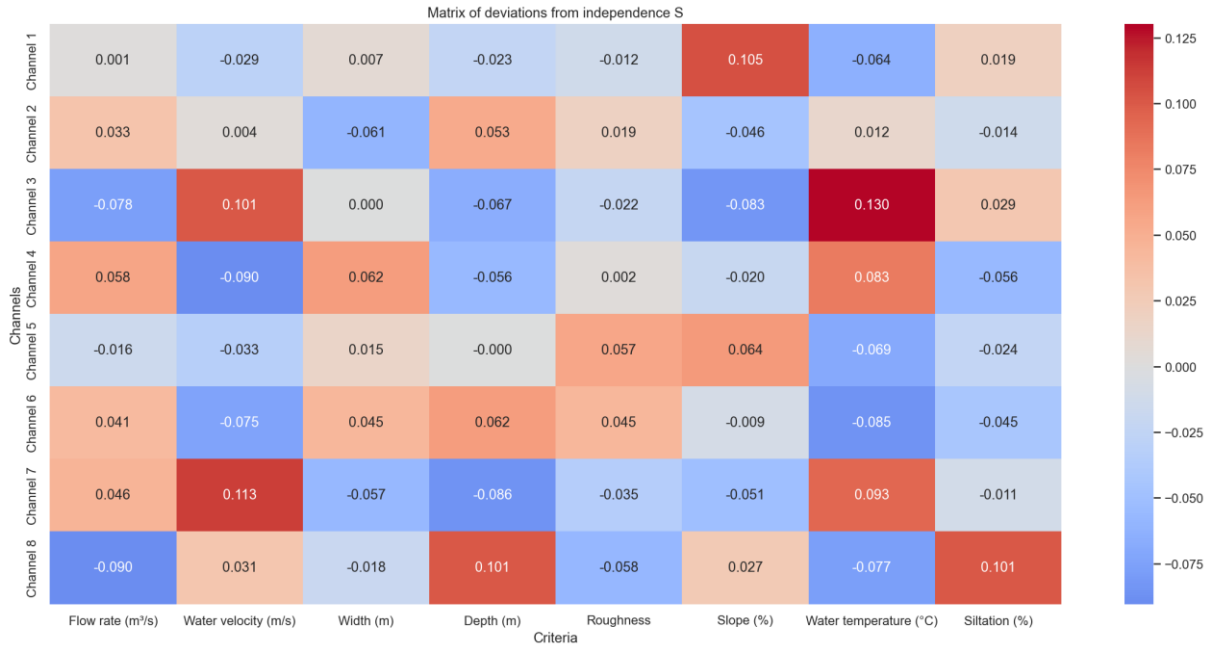


Column Masses (Criteria)

- Flow rate (m³/s) – Mass 0,1361
  Flow rate represents approximately 13.6% of the total mass of the criteria. It therefore plays an important role in the overall structure of the matrix and contributes significantly to the contrasts observed between the channels.
- Water velocity (m/s) – Mass 0,1028
  Water velocity has the lowest mass, around 10.3%. Its overall influence remains moderate compared to the other criteria; it is involved in the analysis, but not decisive.
- Width (m) – Mass 0,1500
  With a mass of 0.15, width is one of the most heavily weighted criteria. It contributes significantly to the differentiation of the channels and plays a major role in the extracted axes.
- Depth (m) – Mass 0,1194
  Depth accounts for nearly 12% of the total mass. It has intermediate importance: not negligible, but slightly less than flow rate or width.

- Roughness – Mass 0,1500
  Roughness shares the highest mass with width (15%). This criterion is therefore central to the analysis and strongly reflects the physical state of the channel bed and its resistance to flow.
- Slope (%) – Mass 0,1056
  Slope has a relatively low mass (10.6%). It is included in the overall description, but its weight remains less than that of the geometric and roughness variables.
- Water temperature (°C) – Mass 0,1167
  Water temperature contributes approximately 11.7%. It plays a complementary role, linked to climatic conditions and the thermal functioning of the channels.
- Siltation (%) – Mass 0,1194
  Siltation has a mass similar to that of the depth (≈11.9%). It contributes significantly to the characterization of canals, in connection with sedimentation processes and the morphological evolution of structures.

Matrix of deviations from independence S



Matrix of Deviations from Independence *S*

This matrix shows, for each channel-criterion pair, the deviation between the observed frequency and the frequency that would be expected if the channels and criteria were independent. Positive values indicate a stronger association than expected, negative values an underrepresentation.

Channel 1: Channel 1 exhibits a slight excess of slope (0.105) and, to a lesser extent, siltation (0.019) relative to independence. Conversely, temperature (-0.064) and velocity (-0.029) are slightly underrepresented, indicating a channel that is rather steep and somewhat silted, but thermally and dynamically less pronounced than average.

Channel 2: Channel 2 stands out with an excess of flow rate (0.033), depth (0.053), and roughness (0.019). The criteria width, slope, and siltation have low values close to zero, indicating a channel with a slightly higher flow rate, greater depth, and a rougher surface than predicted by an independent model.

Channel 3: Channel 3 is strongly associated with velocity (0.101) and temperature (0.130), while flow rate (-0.078), depth (-0.067), and slope (-0.083) are underrepresented. This indicates a relatively fast-flowing and thermally significant channel, but neither very steep nor particularly high-flow or deep.

Channel 4: For channel 4, the main excess values are for width (0.062) and flow rate (0.058), while velocity (-0.090) is significantly less frequent than expected. This channel corresponds to a wide and high-flow section, but with relatively low velocities compared to the other channels.

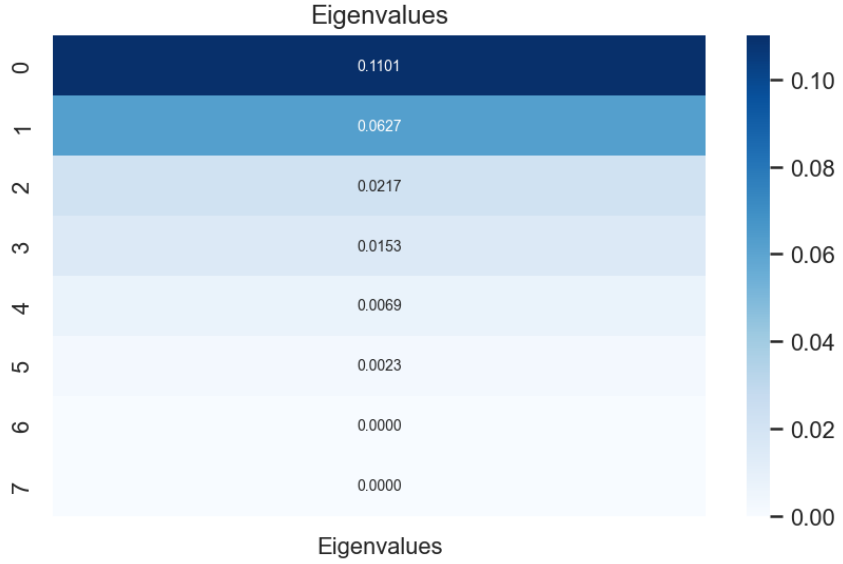
Channel 5: Channel 5 exhibits an overrepresentation of roughness (0.057), combined with a slight excess of width. Conversely, temperature (-0.069) and velocity (-0.033) are underrepresented. This indicates a channel with a relatively rough bed, but with more modest dynamics and thermal influence.

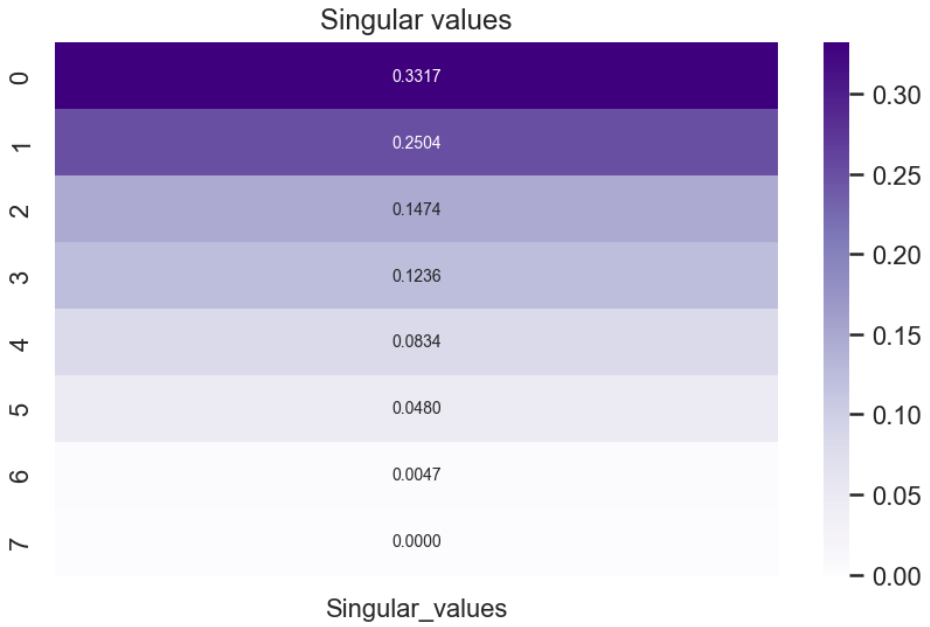
Channel 6: Channel 6 shows a simultaneous excess of flow rate (0.041), width (0.045), depth (0.062), and roughness (0.045), making it a large-sized and morphologically distinctive channel. Temperature (-0.085) and slope (-0.009) are slightly underrepresented.

Channel 7: Channel 7 is characterized by a very high overrepresentation of velocity (0.113) and a positive value for flow rate (0.046), while depth (-0.086), width (-0.057), and slope (-0.051) are less prominent than expected. This corresponds to a relatively narrow and shallow channel, but with a rapid flow.

Channel 8: Finally, channel 8 shows a marked excess of depth (0.101) and siltation (0.101), while flow rate (-0.090) and, to a lesser extent, velocity (-0.031) are underrepresented. This channel therefore appears as a deep and heavily silted section, where the observed flow rates and velocities remain lower than the average for the channels.

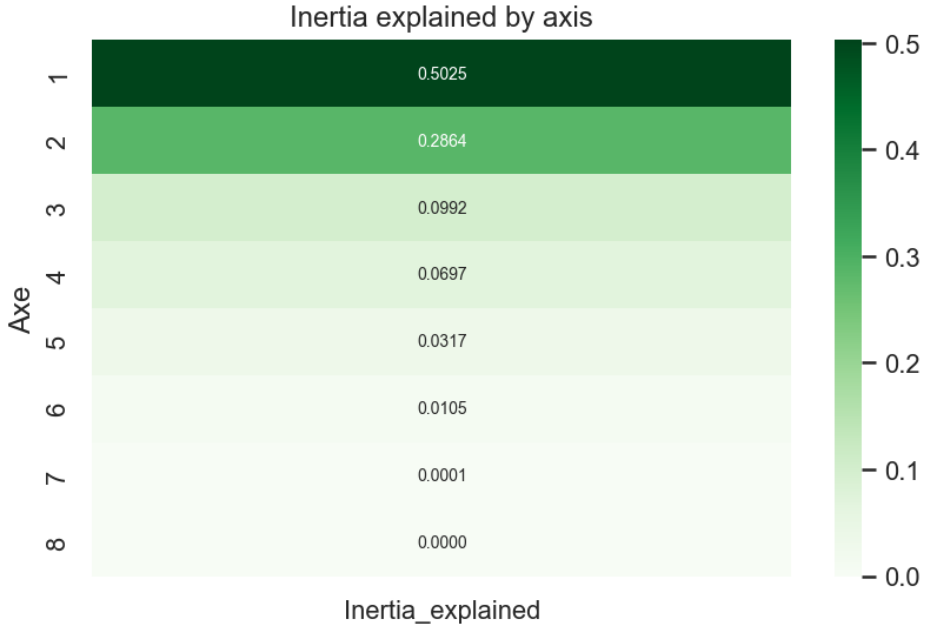
Eigenvalues, Singular values and Inertia explained by axis





Singular values

The decreasing singular values (0.3317; 0.2504; 0.1474; 0.1236; 0.0834; 0.0480; 0.0047; 0.0000) reflect the importance of each factorial axis resulting from the decomposition. The first two values are significantly higher than the following ones, indicating that most of the structure of deviations from independence is concentrated on axes 1 and 2.



Eigenvalues and Explained Inertia

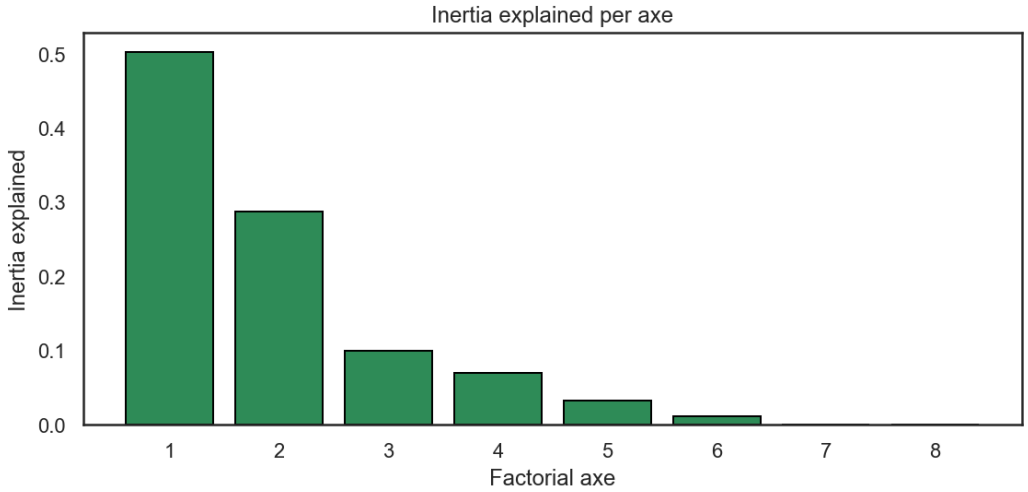
The associated eigenvalues (0.1101; 0.0627; 0.0217; 0.0153; 0.0069; 0.0023; 0.0000; 0.0000) translate into relative inertias:

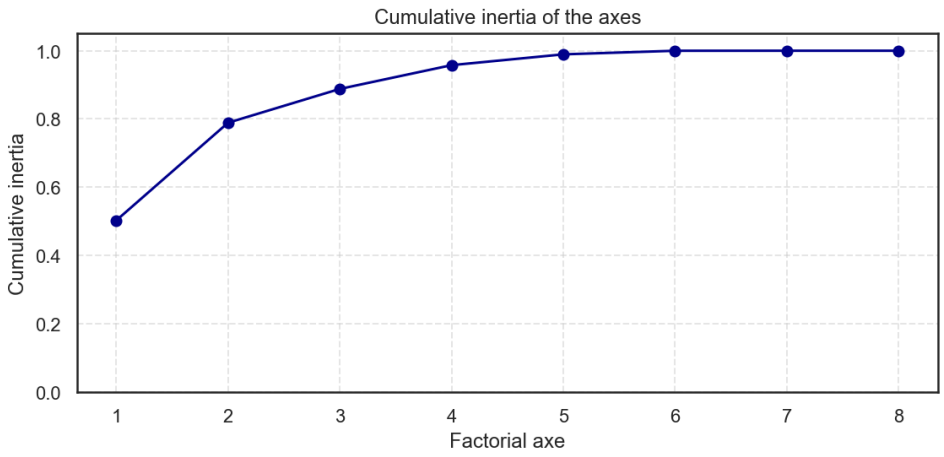
- Axes 1: 0.5025, representing 50.25% of the total inertia.
- Axes 2: 0.2864, representing an additional 28.64%.

- Axes 3 and 4: 9.92% and 6.97%, respectively.
- Axes 5 through 8 each account for less than 3.2% and are therefore considered secondary.

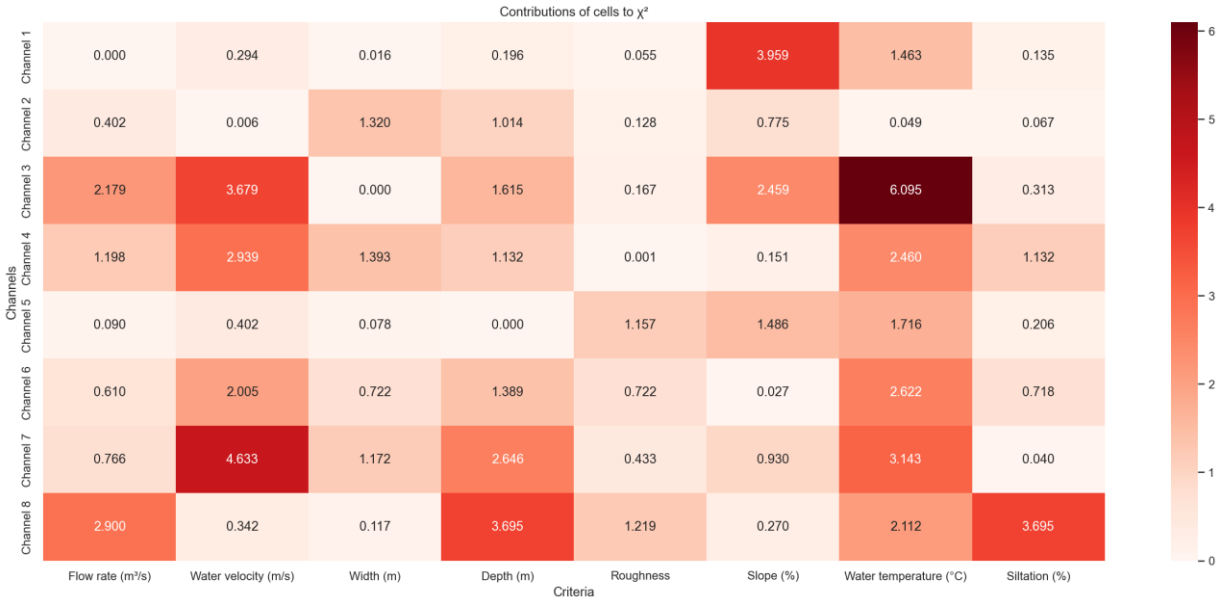
Overall Interpretation (Total Inertia = 0.2190)

The total inertia of 0.2190 represents the total variability of deviations from independence in the contingency matrix. The first two axes alone account for approximately 78.9% of this inertia (50.25% + 28.64%), meaning that a 2D factorial plane is more than sufficient to visualize and interpret the major oppositions between channels and criteria. Axes 3 and 4 provide only a more refined complement (approximately 17% combined), while the last axes primarily capture noise or very specific details.





Khi²



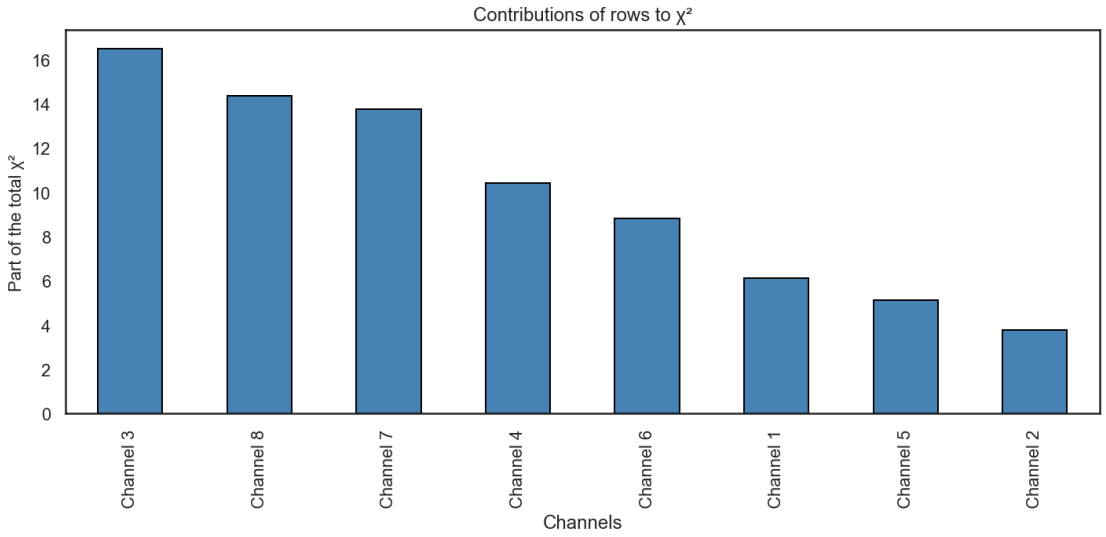
Cell Contributions to the χ^2

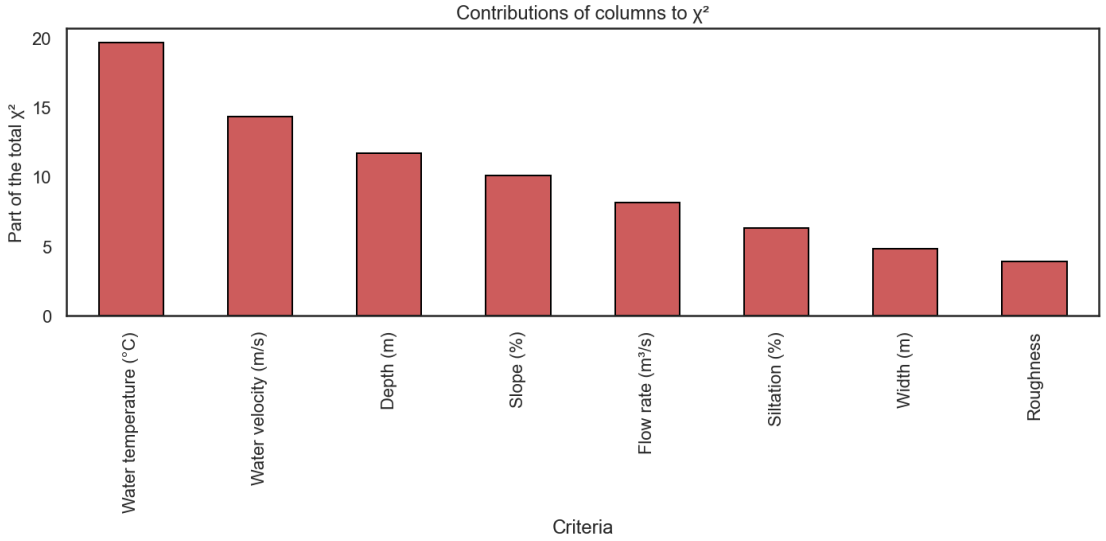
This heatmap shows, for each channel-criterion pair, the cell's contribution to the χ^2 statistic of the correspondence analysis. High values indicate the associations that best explain the overall deviation from independence between channels and criteria.

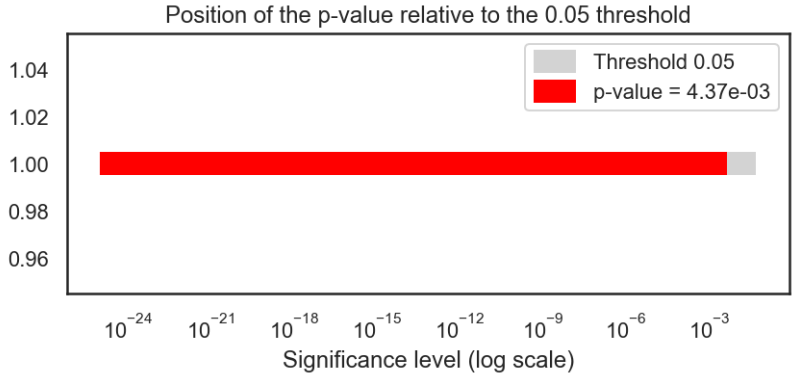
Main Contributions by Channel

- Channel 1: The strongest contributions are for slope (3.959) and, to a lesser extent, temperature (1.463). Channel 1 therefore plays a major role in the association between "steep slope and specific thermal characteristics."
- Channel 2: This channel contributes primarily through width (1.320) and depth (1.014), which shows that it is particularly involved in the geometric contrasts of the network (wider and deeper sections than expected).
- Channel 3: The most significant cells are velocity (3.679), slope (2.459), and especially temperature (6.095). Channel 3 is therefore a key factor in the discrepancies linked to high velocities and unusual thermal conditions.
- Channel 4: Channel 4 is distinguished primarily by its velocity (2.939), width (1.393), and temperature (2.460). It contributes significantly to the "fast and relatively wide channels" contrast with a pronounced thermal signature.
- Channel 5: Its contributions are generally weaker; the most notable are slope (1.486) and temperature (1.716). Channel 5 therefore plays a moderate role in the discrepancies, linked to gradients in slope and temperature.
- Channel 6: The significant contributions are velocity (2.005), depth (1.389), and temperature (2.622). Channel 6 explains a significant portion of the contrasts between deep channels with good flow velocities and specific thermal regimes.

- Channel 7: This channel is strongly characterized by velocity (4.633) and depth (2.646), as well as temperature (3.143). This makes it a key channel for the differences between very fast-flowing, relatively deep, and thermally distinct sections.
- Channel 8: The major contributing factors are discharge (2.900), depth (3.695), and siltation (3.695). Channel 8 is therefore central to the overall variation related to sections that are simultaneously high-flow, very deep, and heavily silted.



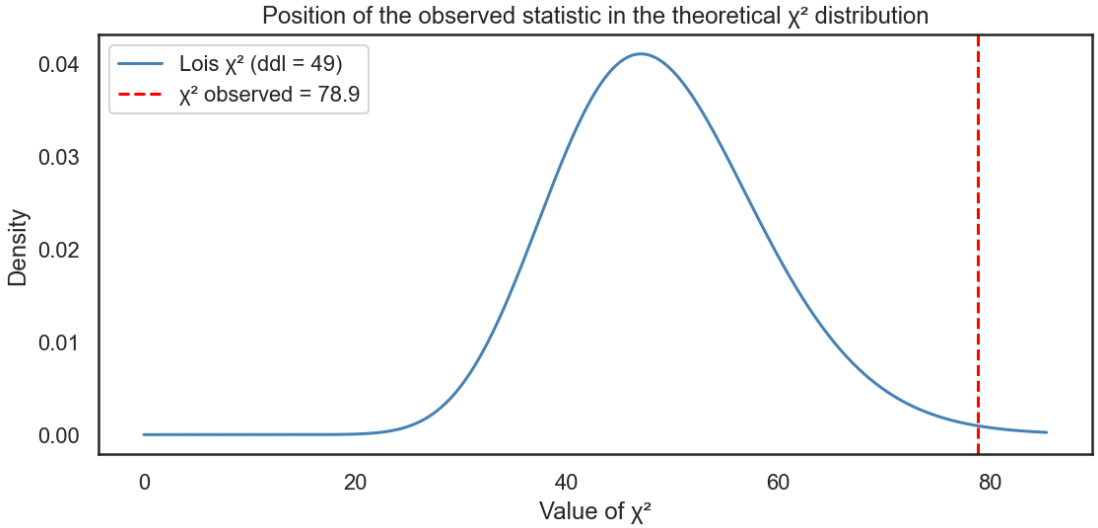




The graph represents the position of the p-value of the χ^2 test relative to the significance level of 0.05, on a logarithmic scale.

- The red bar indicates a p-value of 4.37×10^{-3}, well below the 0.05 threshold (gray area).
- This means that the null hypothesis of independence between channels and criteria is rejected at the 5% level: the discrepancies observed in the table are not due to chance; there is a statistically significant association between certain channels and certain hydraulic parameters (flow rate, velocity, depth, siltation, etc.).

In simpler terms, the structure revealed by the correspondence analysis is significant, and it is legitimate to interpret the factorial axes and the cell contributions.



This graph positions the observed chi-square test statistic relative to the theoretical chi-square distribution with 49 degrees of freedom.

- The blue curve represents the probability density function of the chi-square distribution (49 degrees of freedom), which yields the expected chi-square values under the assumption of independence between channels and criteria.
- The red vertical line indicates the observed chi-square value of 78.9, clearly located in the right tail of the distribution, where probabilities are very low.

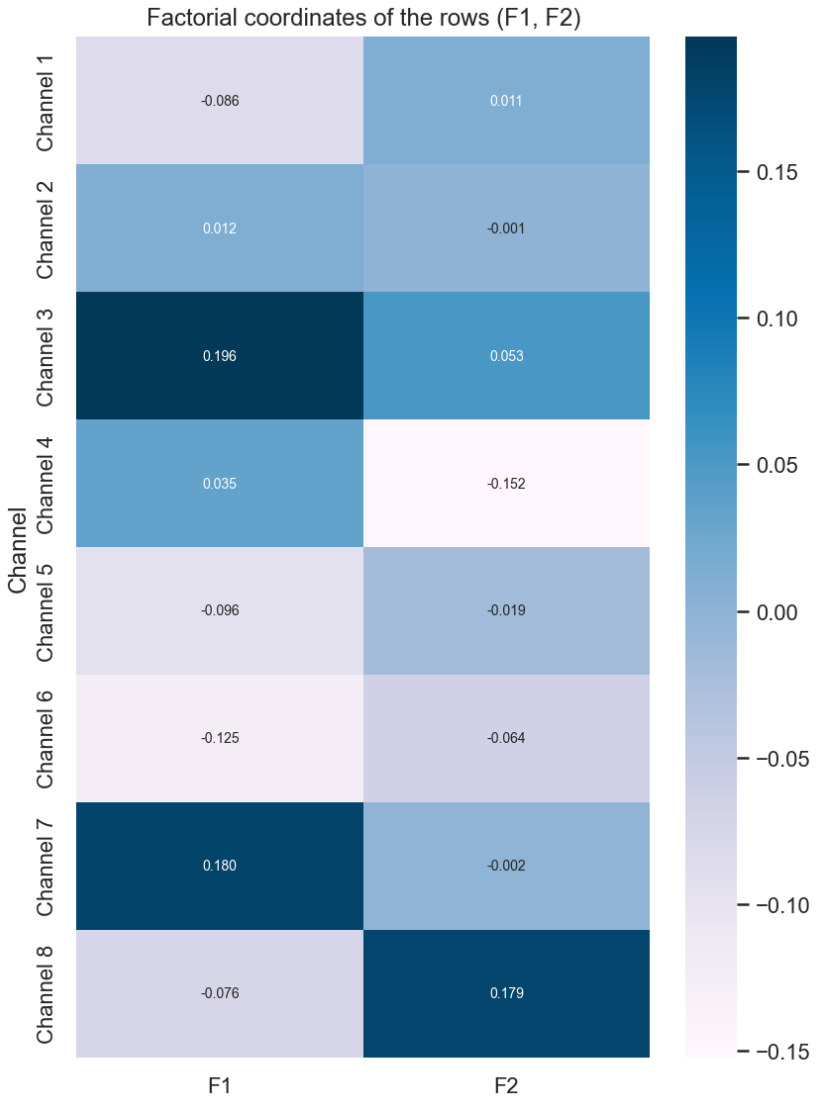
This tail position means that, if independence were true, obtaining a statistic as high as 78.9 would be highly improbable. This confirms that the assumption of independence is rejected: there is indeed a significant relationship between the distribution of criteria (flow rate, velocity, depth, siltation, etc.) and the different channels studied.

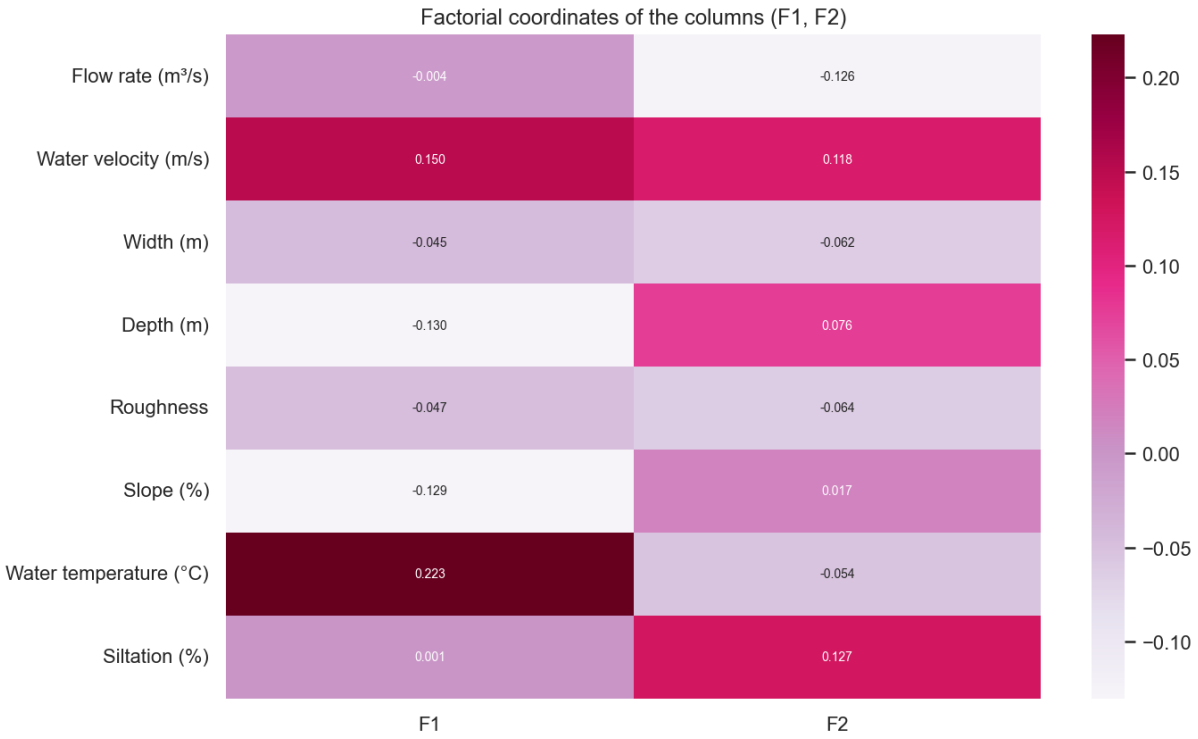
The results of the chi-square test show that the dependence between the channels and the hydraulic criteria is statistically significant. The observed statistic is $\chi^2 = 78.85$, very close whether calculated by correspondence analysis (CA) or by the statistical function, thus validating the consistency of the calculations.

With 49 degrees of freedom, this χ^2 value corresponds to a p-value of 0.00437, significantly lower than the usual threshold of 0.05. This means that the probability of obtaining such a

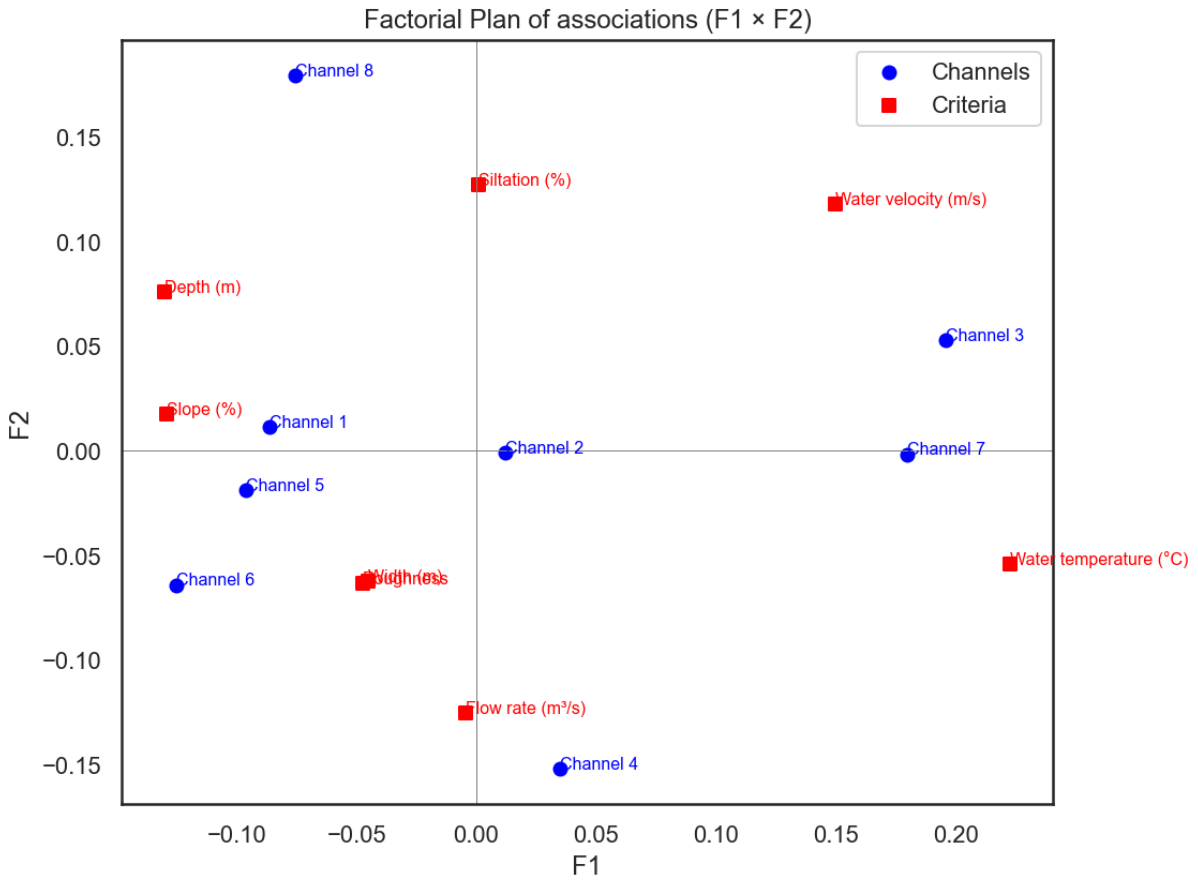
large difference if the channels and criteria were truly independent is very low (less than 0.5%). Therefore, we reject the hypothesis of independence and conclude that there is a significant relationship between the position of the channels and the variables studied (flow rate, velocity, width, depth, roughness, slope, temperature, and siltation).

Factorial coordinates of the rows (channels) and the columns (criteria)





Factorial plane of associations



This factorial plane represents the result of the correspondence analysis (CA) on the first two axes (F1 and F2), positioning both the channels (blue dots) and the hydraulic criteria (red squares).

Axis Direction

- Axis F1 separates, on the right, the variables related to water velocity and temperature (and channels 3 and 7) from those associated with flow rate and width, located on the left.
- Axis F2 contrasts, at the top, the criteria depth and siltation (and channel 8) with those located at the bottom, centered around flow rate and width, as well as channels 4, 5, and 6.

Main Channel-Criteria Associations

- Channel 3 is close to the water velocity vector (m/s): it corresponds to a channel where the velocity is higher than average.
- Channel 7 is also located on the right, in the area influenced by water velocity and temperature, indicating rapid flow in a specific thermal context. • Channel 8 is clearly associated with an upward slope, close to the criteria of depth (m) and siltation (%): it is a deep and heavily silted channel, with a very pronounced morphological feature.
- Channels 4, 5, and 6 are located in the lower part, relatively close to the criteria of flow rate (m^3/s), width (m), and depth (m): they represent relatively wide and flowable sections, but with less siltation.
- Channel 1 is moderately associated with the slope (%), suggesting a channel that is slightly steeper than the others, without extreme characteristics in the other criteria.

In summary, this plan highlights three main profiles:

1. Fast-flowing and thermally marked channels (Channels 3 and 7).
2. A deep and silted channel (Channel 8).
3. Wider, flow-bearing channels (Channels 4–6), with Channel 1 being slightly steeper.

Chapter IV: Cybersecurity

General Context

This dataset represents the monitoring of multiple channels within an industrial/SCADA system, where the goal is to assess the cybersecurity risk level for each channel based on several technical indicators.

Channel Descriptions

- Each row corresponds to a "Channel" (Channel 1 to Channel 20), which can be interpreted as a communication link, a remote station, or a network segment within an industrial architecture.
- The idea is that these channels carry critical commands and data between the control center and field equipment (RTU, PLC, IED, etc.).

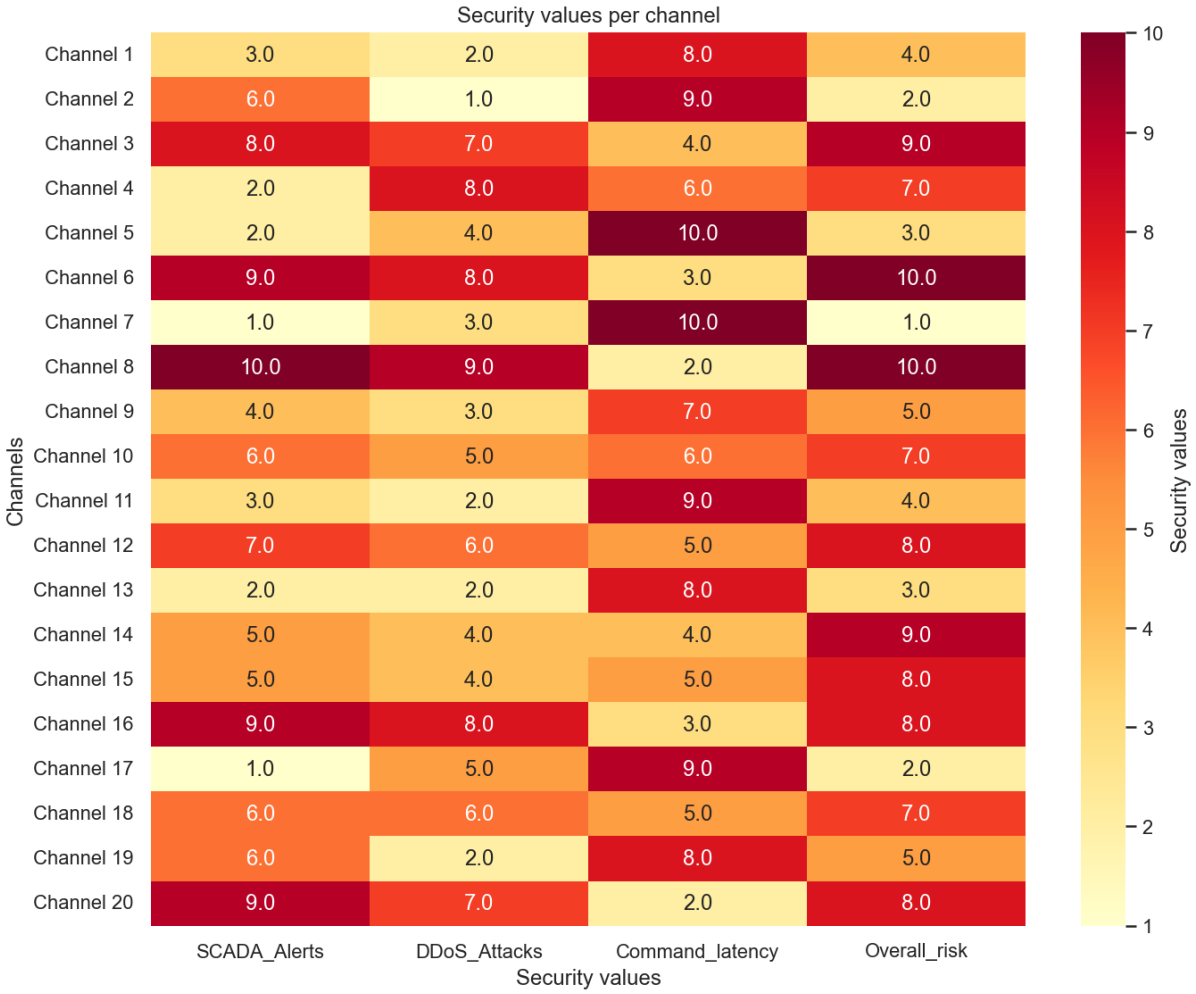
Cybersecurity Variables

- SCADA Alerts: Level (1–10) of alerts generated by the SCADA system or the IDS/IPS on this channel (anomalies, attack signatures, suspicious behavior).
- DDoS Attacks: Intensity or frequency of distributed denial-of-service attacks targeting this channel (network saturation, unavailability).
- Command Latency: Increased latency of commands sent to the field, which may indicate congestion, attack, or failure.
- Overall Risk: Synthetic cyber risk score for the channel, calculated from the other variables (or assigned by an analyst).

Analysis Objective

- Compare channels to identify those with the highest overall risk and understand which factors (alerts, DDoS attacks, latency) contribute most to this risk.
- Use this dataset as a basis for analysis (correlations, clustering, prioritization of security actions, detection of critical channels).

Cyber variables per channel



The four criteria are cybersecurity risk indicators associated with each channel, rated here on a scale of 1 (low) to 10 (very high).

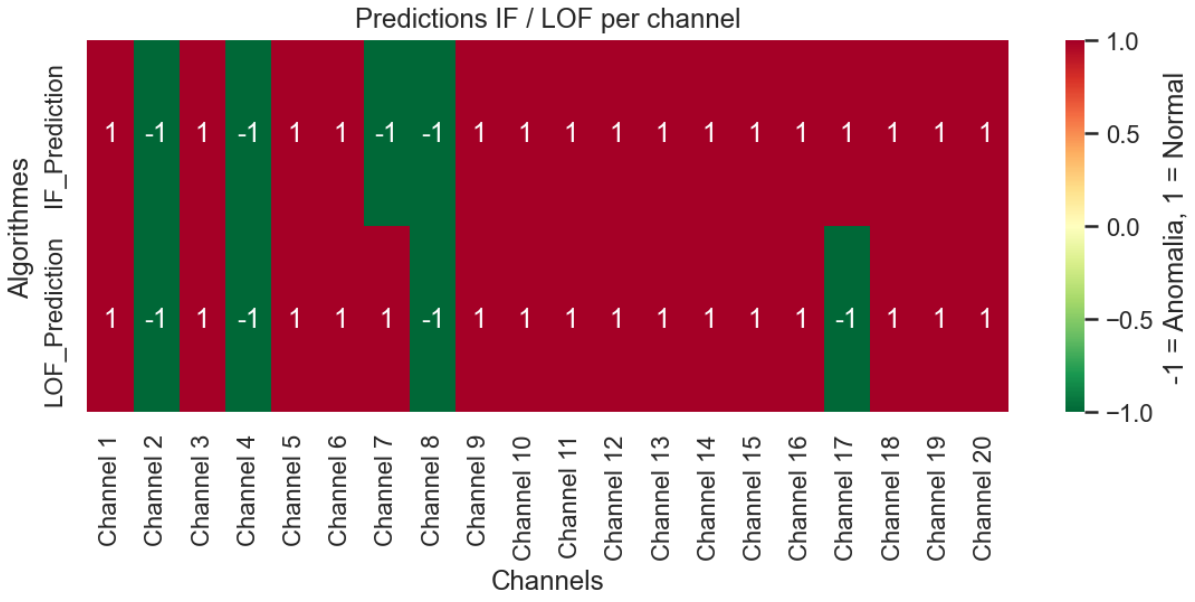
SCADA Alerts: Number or intensity of security alerts detected by the supervisory control and data acquisition (SCADA) system: unauthorized access attempts, communication anomalies, suspicious errors in sensors or PLCs. A high value indicates a heavily used or potentially targeted control environment.

DDoS Attacks: Level of exposure to distributed denial-of-service (DDoS) attacks on the communication networks associated with the channel (SCADA servers, APIs, gateways). A high score reflects spikes in malicious traffic capable of saturating the links and disrupting monitoring.

Command Latency: Degree of latency or degradation in response times when sending commands (valve opening, flow rate settings, etc.). High latency can be linked to network attacks, congestion, or architectural flaws, and increase operational risk by delaying regulatory actions.

Overall Risk: A composite index of overall cyber risk for each channel, combining the three previous dimensions (alerts, attacks, latency) and potentially other factors (vulnerabilities, internet exposure, level of protection). A high value indicates a channel that is a priority in terms of security and monitoring.

Predictions IF & LOF



This heatmap compares the anomaly predictions of two unsupervised algorithms, Isolation Forest (IF) and Local Outlier Factor (LOF), for each of the 20 channels. A value of 1 indicates a "normal channel," and a value of -1 indicates an "anomaly channel."

Isolation Forest (IF_Prediction row)
Isolation Forest (IF) flags two anomalous channels: Channel 4, Channel 8, Channel 2, and Channel 7. Channel 4 is atypical compared to the rest of the channels, and Channel 8 also has a cyber profile that is sufficiently different to be isolated by the algorithm.

Local Outlier Factor (LOF_Prediction row)

LOF flags Channels 2, 4, 8, and 17 as anomalous (-1), while all other channels are considered normal (1). LOF is more sensitive to local densities and therefore detects a group of two atypical channels in the cyber variable space.

Interpretation

By combining the two methods, Channels 4, 2, and 8 are detected by both algorithms, making them the most robust anomaly. Channels 7 and 17 are anomalies specific to a single algorithm and should be considered as cases to monitor.

The other channels are judged normal by both algorithms, which reinforces confidence in their "standard" cybersecurity profile. In practice, Channels 4, 2, and 8 should be prioritized for detailed security analysis, Channels 7 and 17 as cases to monitor, and the others as a reference for normal operation.

Risky channels