WATER QUALITY MEASUREMENTS IN THE BRANTAS, INDONESIA: CONSTRUCTING DIFFERENT PERSPECTIVES WITH PRINCIPAL COMPONENT ANALYSIS

A thesis submitted to the Delft University of Technology in partial fulfillment of the requirements for the degree of

Master of Science in Civil Engineering

by

Tijmen Stefan Willard

June 2022

Tijmen Stefan Willard: *Water Quality Measurements in the Brantas, Indonesia: Constructing Different Perspectives with Principal Component Analysis* (2022)

Data used in this study will not be uploaded with the publicly available these, but a list of files is provided in annex E.

The work in this thesis was made in the:

Sectie Water Resources Engineering
Department of Water Management
Faculty of Civil Engineering & Geosciences
Delft University of Technology

Supervisors: Dr.ir. Maurits W. Ertsen
Dr. Boris van Breukelen
Dr. Saket Pande
ir. Reza Pramana

ABSTRACT

The water quality of the Brantas river in Indonesia is of concern to several agencies on East Java. These agencies all measure its water quality in their own way in terms of locations, rhythms and parameters. The goal of this thesis is to find out if these agencies measure the same and if not, how these measurements differ. From these measurements, perspectives are constructed for each agency with the use of Principal Component Analysis. The agencies investigated are the Dinas Lingkungan Hidup Jawa Timur (DLH Jatim), Belai Besar Wilayah Sungai Brantas (BBWS) and Perum Jasa Tirta I (PJT). As an addition to the PCA, a neural network model is constructed and trained to recognize the measurement agency of a datapoint from the measurement values.

It was found that the all three agencies recognized oxygen as a dominant driver in water quality processes. Secondary processes were mostly driven by rainfall, but the effect of this was seen differently by the agencies. DLH Jatim distinguishes surface waste runoff separately from rainfall, while Perum Jasa Tirta I will see them as inherently connected. BBWS will not recognize the surface waste runoff process as a significant factor in the water quality. These differences found in the representation of core processes in the Brantas outline how different agencies can have a different perspective on water quality. This was further underlined by the conclusions from the neural network analysis. Here it was found that the author could be recognized from the measurement values alone on 88% of agency data.

ACKNOWLEDGEMENTS

This thesis will be handed in as the end product to fulfill my MSc in Civil Engineering. It has been quite the adventure and I want to thank some people for helping me along the way. First of all I want to thank the members of my thesis committee, ir. Reza Pramana, Dr. Boris van Breukelen and Dr. Saket Pande, for their valuable advise and insights. Next I want to thank the students and teachers at the section of Water Resources Management for their social support and for facilitating me. Most importantly though, I want to thank my supervisor, Dr. ir. Maurits Ertsen, whose support and advice was of immeasurable value to me and was instrumental in helping me finish the thesis. I cannot express my gratitude enough for that.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

As the second largest river basin on the island of Java, the Brantas river basin provides essential services to some 17 million Indonesians. This varies from water for household and industrial use, to irrigation and power generation (Valiant, 2013). The Brantas basin has experienced intense industrialization, agricultural development and rapid population growth over the past decades; changes that bring with it a vast amount of challenges (Dueñas, 2008). Maybe the most severe of those is the increase in pollution from these activities. Waste water from agriculture, households and industry flows directly into the river, most of it untreated. In addition to that, lack of proper waste management facilities has created a situation in which a lot of solid waste, most of which plastics, is dumped into the river as well (Visser, 2019). It should be clear that this pollution is seriously problematic for a river that is also a precious resource.

A large amount of stakeholders have an interest in good water quality management. There are private actors such as industries and businesses or NGO's and environmental organizations, but maybe more importantly there are the people in the basin who depend on the Brantas for their livelihoods for drinking water, irrigation and fishery. Just like other countries, Indonesia has a large bureaucratic organization that makes policy for water quality management. There are the central government's Ministry for the Environment and the Ministry for Public Works, there are several Basin Management Agencies, there is the Provincial Government of East Java and the numerous city councils and local regencies that exist in the Brantas basin (Houser, 2021a). In theory, these institutions should cooperate in deciding common goals for water quality management, but in practice this coordination can be problematic. Many agencies find themselves in a fragmented landscape with often overlapping activities and policy mandates (Houser, 2021a). That is in itself not so strange as countries such as the Netherlands struggle too with good inter-agency coordination and cooperation with regards to water quality management (Junier, 2017). That, however, does not diminish the necessity of coordination for effective water quality policy.

One such area with an overlap of activities is monitoring of water quality. At least three agencies have been identified with large scale water quality measurement campaigns in the Brantas: a public company and two government agencies, one on the provincial level and the other on the basin level. They will be further introduced in section 1.1. This overlap of research actions does provide an interesting opportunity though, because this knowledge of three different measurement sets from three agencies can be used in different ways. It can be assumed that all data is complementary or that they are just duplicating each others efforts. It can also be that one agency is 'obviousy' right and therefore others must be wrong. All of these ideas assume that there is a certain state of the river's water quality and that it can be measured accurately. While not untrue, a river is a very complicated thing and small choices like when or where exactly to take a sample can make a big impact on the values that you are going to find. So even when different agencies report different values, they may simply both be true. Small and unconscious choices may seemingly have little impact, but over an entire measurement campaign, they can

add up. An agency can start to construct a story with the measurements it has done, but this story may not be the same as the other agency that is measuring the same river, but different values. This research is interested in these differences in the measurements and the different stories these may create. The central questions are as follows:

- How do measurements in the same river, differ across different agencies?
- How do the measurements shape a perspective of the river?
- What are these perspectives?

Before I dive deeper into this, let me first introduce the actors.

1.1 AGENCIES

1.1.1 Dinas Lingkungan Hidup Provinsi Jawa Timur

Dinas Lingkungan Hidup Provinsi Jawa Timur, which in English can be interpreted as Environmental Protection Agency of East Java or EPA, as which I will refer to it, is a provincial government agency under the governor of East Java Province and is supported by the Ministry of Environment and Forestry of the Republic of Indonesia (KLHK). Related to water quality, its main responsibilities are to ensure a decent water quality of the Brantas river and to enforce regulations on discharge of pollution by industries (Visser, 2019, Annex E.5). In its 2019-2024 five year strategic plan (Houser, 2021b), the EPA outlined three main goals:

- Control the pollution in the river basin by improving the percentage of businesses with environmental assessment documents.
- Increase the coverage of garbage and hazardous (B3) waste management services.
- Increase the water quality of river water on BOD and COD parameters.

The EPA also has the authority to enforce pollution standards on industries by handing out fines (Visser, 2019, Annex E.5), but current law enforcement is rather weak (Houser, 2021b; Dueñas, 2008). It seems obvious that failing to enforce industry compliance with environmental standards would result in more non-compliance and thus increased pollution and a deteriorating water quality. Hence, it is remarkable that the improvement of enforcement is not a key focus point for the EPA. Enforcement of environmental standards is not trivial though and multiple reasons can be attributed to its weak state. Most important are the lack of qualified staff to serve as environmental enforcement officers and low budget support (Houser, 2021b).

A goal that is defined though is to increase the percentage of industries with environmental assessment reports. One such assessment report is the PROPER program, a national government program for which the EPA is responsible in East Java. This is a voluntary program where businesses and industries report their pollutant discharge, which in turn is verified by the EPA. At the end of the year, environmental labels are handed out in five different classes: black, red, blue, green and gold, with black being the worst and gold being the best. The incentive to improve one's label is mostly for reputation. The program has received praise from the World Bank (Makarim et al., 1995) for its ability to implement a large environmental monitoring program with very limited resources and information. Additional arguments from Heyes (2000)

also suggest that the PROPER program can be an effective tool to increase compliance with environmental standards. However, Heyes (2000) also stresses that more traditional environmental enforcement methods should not be overlooked and that reputational incentives alone will likely not be enough to convince every actor to comply.

1.1.2 Balai Besar Wilayah Sungai Brantas

Balai Besar Wilayah Sungai Brantas (BBWS), or Grand Office of the Brantas River Basin in English, is a river basin authority that falls under the responsibility of the General Directorate of Water Resources (DJ SDA), a government agency of the Ministry for Public Works and Public Housing (KPUPR). BBWS' main task is to provide policy on water quality and quantity (Visser, 2019). According to its own website (BBWS, 2020, Organization Profile), its mission is as follows:

> *"The Grand Office of the River Basin has the task of carrying out the management of water resources in the river area which includes program preparation, implementation of construction, operation and maintenance in the context of conserving and utilizing water resources and controlling the destructive power of water in rivers, beaches, dams, lakes, lakes, reservoirs. and other water reservoirs, irrigation, swamps, ponds, groundwater, and raw water as well as the management of urban main drainage."*

To summarize, BBWS is mostly concerned with policy on water resources. This includes both water quantity and quality. Day to day management and operations technically fall under the responsibility of BBWS, but is mostly done by Perum Jasa Tirta I (see below; Visser, 2019). For water quantity, the goals seem quite clear: prevent flooding and optimize the use of reservoirs. For water quality on the other hand, goals are much less clear. BBWS has a mandate to outline policy, but implementation is left to other agencies, both for water quality and quantity. As for now, most set policy concerns the monitoring of water quality (Houser, 2020). As for actual management of the river itself, the target is set as "improve water quality throughout the Brantas River". This is a very broad goal, that lacks clear indicators. This observation can be made in general for the policy of BBWS regarding water quality. BBWS has a clear concern about the water quality of the Brantas, but clear goals with regards to this quality are not yet available. At the moment, most efforts go to monitoring and public information campaigns (Houser, 2021a). However, with its responsibilities for the operation of major river infrastructure and urban drainage systems, BBWS is in a good position to lay out more interventionist policies.

1.1.3 Perum Jasa Tirta I

Perum Jasa Tirta I, or Water Service Company 1, is a state owned company set up to manage water resources in the Brantas and Bengawan Solo rivers, the two largest rivers on Java. Together with Perum Jasa Tirta II, who is in charge of rivers on West-Java, PJT1 was established on 13 October 1999 by central government decree (PP No. 93 and PP No. 94). At the time of their establishment, the two PJT's were very much still an experiment to privatize water resources management (Dueñas, 2008). PJT1 (to be further referred to as just PJT) has since then grown, as it now also provides services in the Serayu, Bogowonto and Jratunseluna basins on Java and the Toba Asahan basin on Sumatra (PJT1, 2022, Homepage).

As a water service company, its main responsibility is water allocation and operation and maintenance of water infrastructure (Visser, 2019). Its core business is the operation of reservoirs (see relation with BBWS) for power generation and supply of 'raw' water to industries and drinking water company PDAM (PJT1, 2022, segmen jasa air). With this comes also its main interest in water quality, as it is dependent on the Brantas for the source of this untreated water. PJT thus needs to ensure that the water quality is decent enough for industries and drinking water companies to use. It is authorized to make technical and operation policy decisions, but cannot decide on policies with regards to basin planning or the development of new infrastructure (Dueñas, 2008). This creates a situation where PJT has a large responsibility with regard to the water quality of the Brantas, but little effective policy power to change it. This makes effective cooperation and communication with other authorities essential to effectively conduct its business.

1.2 PROBLEM STATEMENT

Now that the three main actors have been introduced, a closer look can be taken to their water quality monitoring. From the data they provided, a map of the Brantas was drawn up that shows their measurement locations.

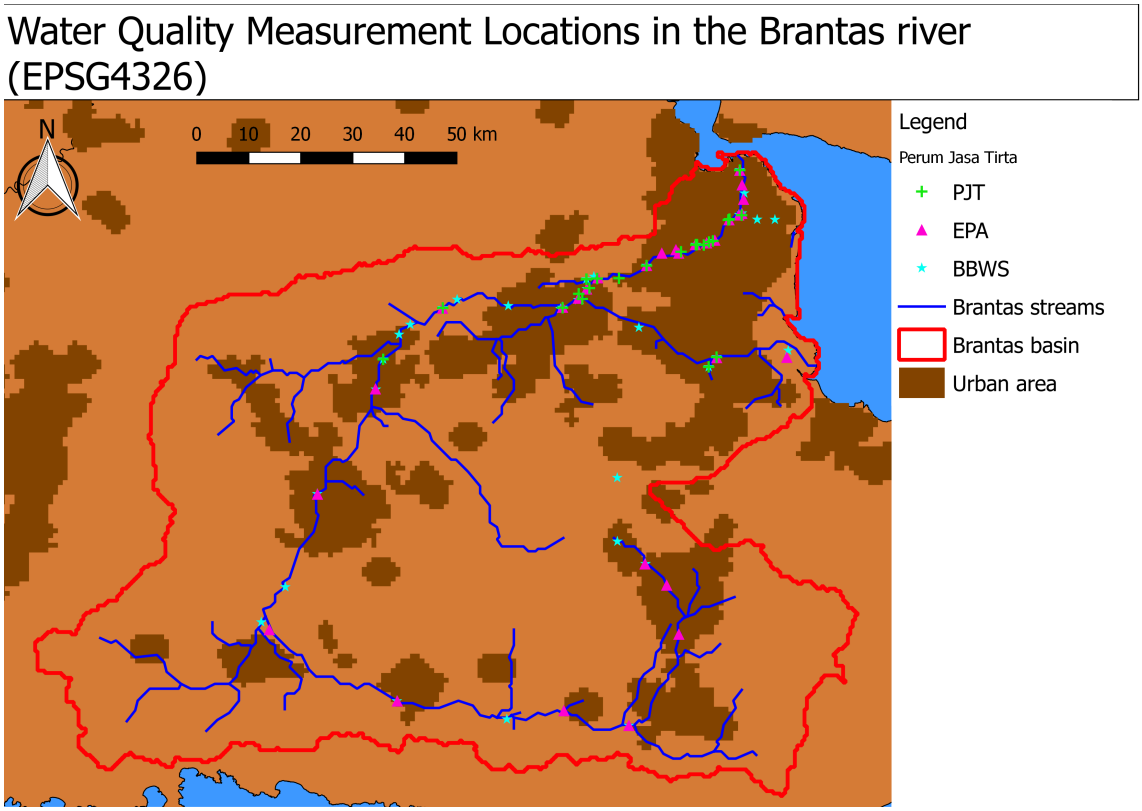

Figure 1.1: Map with measurement locations of the tree main agencies in the Brantas river basin.

One of the first things to notice is that on this scale it is quite cluttered, especially in the northern section. There are a lot of locations so close together that the markers overlap. This occurs mostly in the Surabaya branch of the Brantas, but at other locations as well. Most of the locations are measured by just a single agency, but there are plenty with two or more. As for the patterns of each agency, PJT has many locations in the Surabaya branch, some in the Porong branch and some in the lower sections of the Brantas. It does not do any measurements in more upstream sections. The measured sections are not just downstream, but also heavily urbanized, featuring most of the industry in the basin, so higher pollution loads should be expected here. The EPA has measurement locations spread out across the Brantas. Upper, middle and lower sections are all covered, but interestingly, locations are always in or near an urban area. This does not necessarily say much though, because there are urban areas all along the Brantas. The larger urban areas of Malang and Surabaya also get a bit more attention than other areas. BBWS has its locations fairly spread out over the Brantas. However, although poorly visible on the map, many locations are at the outflow points of some of the tributaries rather than in the main branch. There remain plenty of measuring points in the main branch, but it is not as many as the map would suggest. BBWS also seems much less focused on urban areas and more on the major coalescence and bifurcation points of the major branches.

There is also the matter of frequency, as the agencies do not measure at the same times. The EPA and PJT take monthly measurements, whereas BBWS takes its measurements quarterly. Monthly will obviously give you more information, but that does not mean the coverage is that much better as there is still about 29 other days and 23 other hours on that day to collect information from. River water quality is complex enough to a point where it is not really possible to capture all information possible, so any measurements taken are by definition a simplification. Even with the EPA and PJT both measuring monthly, they do not necessarily take measurements on the same day or time, so both will likely find somewhat different results.

Lastly, there is the question of what substances the agencies measure. Water quality is a generic term, not a specific physical quantity that can be measured directly. Instead, it is explained by a collection of physical quantities or parameters that can be measured. Which parameters to choose from is not necessarily a trivial choice. Different priorities in policy and budget can shape the focus of a measurement campaign as well as the methods chosen to measure certain parameters.

All of these choices in where, when, what and how to measure can and will influence the results that are found. It does not mean they are bad results, but it will influence the perspective those results provide of the water quality of the river. In this research, an attempt is made to find these perspectives or these three agencies in the situation of the Brantas.

The next chapter 2 will go over how the agency data was dissected and treated to extract these perspectives. In the results chapter 3, an overview of the data itself is presented as well as the results of the analyses used. Implications and limitations of the analyses results for the agency perspectives are discussed in chapter 4. Lastly, chapter 5 will present the agency perspectives, together with some recommendations for future monitoring.

2 METHODOLOGY

The methodology will be built up as follows. First, a section on data treatment is presented. Data treatment refers to all the actions done to create a workable dataset from the collected agency data. In this section will also be an explanation of which parameters to move forward with. In the next section, the meaning of these parameters in terms of water quality is briefly explained. In the third section, the main analytic method used, principal component analysis, is explained. This section will also go into how the principal component analysis is applied to the data. The final section will outline the methodology of data classification with neural networks, including why this approach is relevant in the context of this study.

2.1 DATA TREATMENT

Data treatment is subdivided in three steps:

1. Treating loose datafiles and fitting them together in a composite dataset
2. Matching and harmonizing different location titles and creating a consistent system for location information
3. selecting usable data for further analysis

2.1.1 Agency files

Water quality data was requested from all three agencies through the contacts made in the Aksi Brantas project. Communication went through ir. Reza Pramana from whom I received the agency files. These files do not include all water quality data from the Brantas or perhaps even all measuring agencies, but only the ones made available to us. These data sets proved to be useful enough though and the interest of the research is directed towards these agencies, so no effort was taken to include additional data.

In order to fairly compare the agencies, the obtained data needed to be formatted in the same way. The format in which the data was presented was not sufficient to be used in analysis directly and thus some degree of treatment was necessary. This treatment was different for every agency and sometimes even for different files within the files of one agency. In the next sections, there will be a more detailed description of the changes made to the different data files. The common end goal was to create a single datafile which could easily be processed in further analyses.

The final datafile created will be referred to as the composite dataset. This composite dataset contains 5240 datapoints, of which 1151 from the EPA, 1430 from BBWS and 2659 from PJT. A datapoint is a specific time and location with one measurement value for each parameter. A datapoint does contain information for multiple parameters.

EPA

The data from the EPA consists of 5 files for years 2013-2017. The files from 2014 and 2015 were identical. As 2014 was referenced more often in the file than 2015, it is assumed that this data is from 2014. No data from 2015 is thus processed. Data from the EPA is presented monthly, so for every month with data, the file will contain a separate sheet. The data was imported into Python and aggregated from multiple sheets. The sheets within a file tended to have a consistent format, but the files from each year were slightly different, so this process required a mix between manual and automatic imports. Additionally, some editing had to be done to the datfiles in order to make them machine readable. An overview of these files and edits is presented below.

Notes:

- 2013: Only TSS, DO, BOD, COD, TP, Fecal Coli and Total Coli measured. 30 Locations.
- 2014/15: 30 Locations. Measured is: Temperature, pH, EC, TDS, TSS, DO, BOD, COD, NO_3, NH_3, free chlorine, TP, Oil and Grease, Detergents, Fecal and Total Coli, Cyanide and hydrogen sulffide. NO_2 and Phenol is noted, but never assigned a value.
- 2016: 30 Locations noted down, only 28 measured. Measured is: Temperature, pH, TDS, TSS, EC, DO, BOD, COD, TP, NO_3, NH_3, NH_4, various metals (Co, Cd, Cr, Cu, Fe, Pb, Mn, Zn), cyanide, fluoride, NO_2, SO_4, free chlorine, hydrogen sulfide, Fecal and Total Coliform, Oil and Grease, Detergents and Phenol.
- 2017: 24 Locations, measured only in 5 months (Jan, Feb, Oct, Nov, Dec). Measured parameters are identical to 2016 with the exception that NH_4 is not measured anymore.
- It occurred multiple times that a measurement was assigned as lower than a certain value. This is converted to that value instead.

BBWS

The data from BBWS consists of 11 files that range annually from 2009 to 2019. BBWS measures quarterly, so at most 4 different sheets with data per file were provided. This is why the EPA and BBWS have a fairly similar number of datapoints despite BBWS covering a much longer timeframe. The exact date of measurement is often missing from the BBWS data and sometimes the month of measurement is unknown as well. Another issue, similar to the EPA data, is that the template in which the data is presented changes constantly, making the files hard to work with. This sometimes even occurs when going from one sheet to the other within a single file. More detailed information is presented below.

Notes:

- In some pages of the documents, measurements were described as non-detectable. Here they have been assumed as a value of zero. In other pages, there are instead dashes. As

it is not clear whether the parameter was non-detectable or not measured, these have been converted to NaN (Not a Number) values instead. An exception was made for months that contained data at other locations for the same parameter, but not for that specific one. Here it is deemed more likely that the parameter was measured, but the contaminant was not detected instead of assuming no measurement was made at all. Henceforth, empty rows are assumed to contain no data, while single empty fields are assumed to contain zero-data.

- Missing date information of which the month is known are set to the first of the month.
- 2009: 8-12 locations measured per month. None are measured twice in a quarter. 17 parameters noted down: Temperature, pH, DO, BOD, COD, TSS, NO_3, NO_2, NH_3, TP, Detergents, Oil and Grease, Phenol, Cu, Cr, Fecal Coli and Total Coli.
- 2010: only has one data sheet. This contains averages of all data points measured over the entire year. Individual values cannot be read by a computer, so the file is declared unworkable.
- 2011: no datetime provided, only the quarter. 36 locations with 11 parameters: Temperature, TSS, TDS, Turbidity, EC, pH, DO, BOD, COD, Fecal Coli and Total Coli.
- 2012: 36 locations quarterly, months have been noted down. Same parameters as mentioned under 2011 with addition of NH_3, NO_3, NO_2, TP, Cu and Cr.
- 2013: Format identical to 2012
- 2014: Format identical to 2012
- 2015: Same parameters as 2012, though Cu, Cr, Fecal Coli and Total Coli had no values. Only quarter was mentioned, but not specific month, so first of quarter assumed. Third quarter has only 19 of 36 locations, quarter four has the remaining 17, but not the other 19.
- 2016: Structure similar to 2012-2014. 13 parameters which are: Temperature, Turbidity, TSS, TDS, EC, pH, DO, BOD, COD, NH_3, NO_3, NO_2 and TP. Number of locations is either 12, 24 or 36, mostly 24, but not consistently.
- 2016: Months filled as first of quarter. Low COD values (many zeros where those would not be expected) and many values lower than the BOD value of the same measurement point.
- 2016: October and November have misnaming of some locations. Compared to the usual order, locations appear in 5 have seemingly shifted to the left, while 3 locations have dropped out and 3 others are observed twice, though different values are recorded. It was assumed the author made a mistake in copying column names, so the stated locations have been replaced by the column names in the order that they appear on other tabs and in other files.
- 2017: Quarterly data with the 13 parameters listed under 2016. 36 locations covered with the exception of the fourth quarter which had 24 locations covered.
- 2018: Quarterly data on the 13 parameters listed under 2016. 36 total locations covered, but quarters 1, 2 and 4 only feature 24.
- 2019: 36 locations covered, 13 parameters as listed under 2016. Quarterly data, but first quarter is missing.

PJT

The PJT data consists of 10 annual files of years 2010-2019. Every file contains a new sheet for every month with recorded dates. Because PJT has different measurement frequencies per location, the number of measured locations varies by month. There are 20 different locations of which 13 are recorded every month, 4 are recorded twice a month and 3 are recorded every 3 months, which is usually during the first month of the quarter. The consistency between datafiles is decent, so some automation in imports is possible.

PJT measures a large number of parameters. All parameters mentioned are: Temperature, pH, DO, Secchi depth, Turbidity, Discharge, BOD, COD, $KMnO_4$, TSS, TDS, Fluorine, Chloride, NO_3, SO_4, Na, Ka, Mg, Ca, Alkalinity, Acidity, NO_2, NH_3, Kjeldhal-N, Total PO_4, TP, Dissolved PO_4, Boron, Sulfide, Hydrogen Sulfide, Cyanide, Chlorine, Phenol, Detergents, Oil and Grease, Cr6, Ag, As, Ba, Cd, Cu, Cr, Fe, Hg, Mn, Ni, Pb, Se, Zn, Al, Co, Total Coli and Fecal Coli.

The actual amount of measured parameters is considerably lower though, as many of the data columns remain empty. There is also a large group of parameters that do have some measurements taken, but are far from complete. The most recorded parameters for PJT are water temperature, EC, DO and COD with 2659 times measured, followed by pH with 2658 times and TSS with 2653 times. Furthermore relatively complete are Turbidity, BOD and NO_3 and NO_2 (all measured 2635 times), Detergent (2612 times), NH_3 (2574 times), Oil and Grease (2551 times) and Phenol (2523 times). Less complete, but still rather decent are TP (1845 times), Fecal Coli (1285 times), Total Coli (1276 times), Copper (1257 times) and Chromium (1253 times). All other parameters were measured less than 1000 times.

There were some other issues with the PJT data that need mentioning. The first and probably most annoying one was that the Excel sheets of the 2019 data had saved a lot of the measurement values as a date rather than a numeric value. For example, a water temperature was saved as 28th of May 2020. If Excel's tools were used to automatically convert to numeric, this would be converted to 43979, which is obviously not a real recorded temperature. The 28th of May was much more likely to be 28.5 °C, which was also what was visually present in the cell. This must have been converted and saved as a date in post processing even though it clearly should not be. This error occurred quite frequently across this file in the sheets from April to December and on the parameters of: water temperature, pH, Dissolved Oxygen, Turbidity, Biochemical oxygen demand and Chemical Oxygen Demand. In the end, the dates were repaired manually from the visual number, which was a date according to Excel, to the actual number.

Another issue with the 2019 datafile occurred on the nitrate measurements. Many of the values were in the thousands of milligrams per liter, an unrealistically high concentration. Interestingly, there were no values between 1 and 1000 mg/L, but there were values below 1 mg/L with three decimals. This created the suspicion that something must have gone wrong with the decimal separators. Because this also occurred in the sheets from April to December, the sum of reasons required this to be corrected and hence all these $>$1000 values were divided by 1000 to bring the values to a more reasonable range.

The last issue worth mentioning were two locations that were only indicated with a number. In the files of 2010 to 2014, two locations were named 4110 and 4500. These were not numbers present on the sheet of locations or any other piece of data from PJT with location information. Therefore some investigation was required to find out which actual locations these represent. This occurred only on the November sheets of 2010-2014 and the December sheets of 2012-2014,

specifically in rows 18 and 19. In sheets of other months, as well as files of other years, these rows were usually filled with Jembatan Ciro and D/S Intake K. Pelayaran. Therefore they were changed to represent these two locations.

The only exception in the other sheets was in December of 2011 where the spots were filled as Jembatan Sepanjang and Bend. Gunungsari. This was odd in and of itself, because these locations were already on that same sheet just two rows above. The date indicated that these were measured with just 4 days in between which created extra suspicion as other locations that were actually measured twice a month were measured with a near perfect two week interval. Therefore, it was accepted that these two locations were duplicated incorrectly and Jembatan Ciro and D/S Intake K. Pelayaran were the correct locations. These were thus also replaced.

2.1.2 Location matching

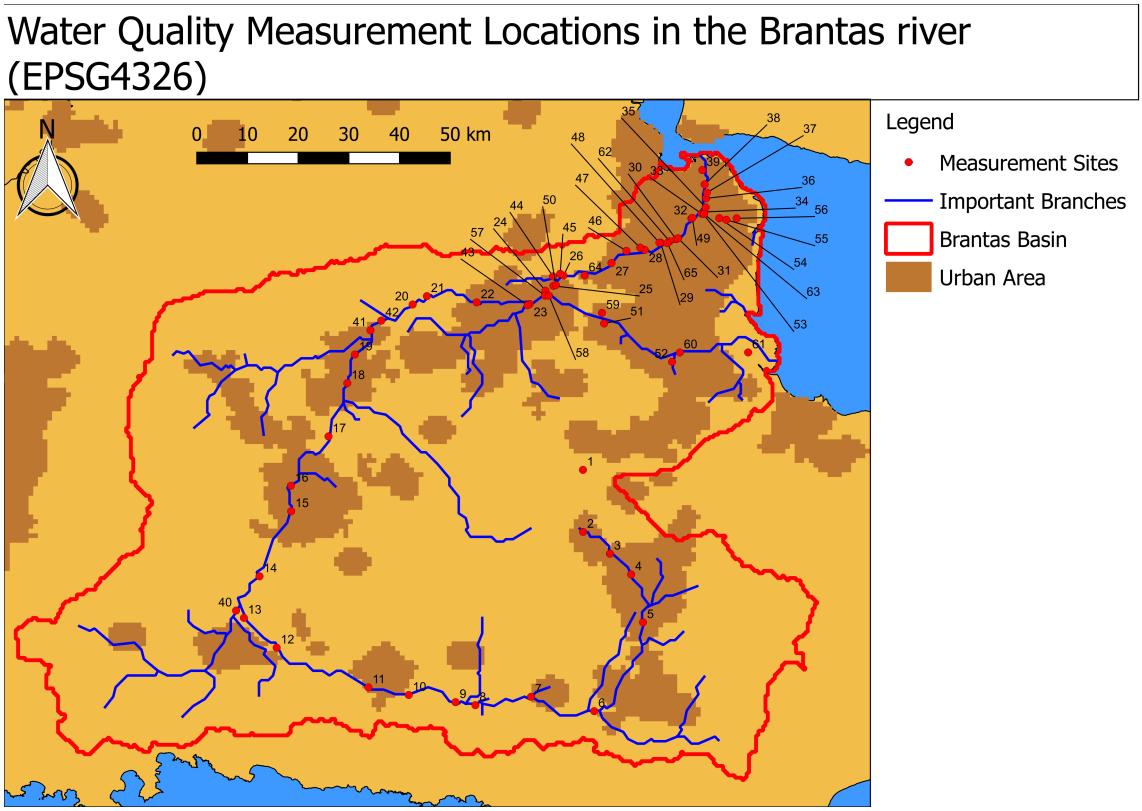

Figure 2.1: Measurement Locations in the Brantas. Numbers correspond to locations as seen in tables of this annex and other figures such as in the diagrams of annex B.

Each agency had its own system of location recording and naming. If all agencies were to measure different locations, this would not be much of an issue. It would require at most some harmonization in the format of naming, e.g. abbreviating frequent words like 'Jembatan' (bridge) to 'J'. As was already seen in figure 1.1, the agencies sometimes measure in the same place. For

these locations, it makes sense to give them the same name. Doing so was not a trivial exercise though.

Most names either referenced the name of a bridge, ferry or outflow point of a sidebranch. That does not mean that these names were always easily recognized. Muara Kali Kwangen and Jembatan Perning for example, refer to the same place, but one would not notice this from just the name. The first names it the outflow of the Kwangen river branch, the other to the bridge going over that. Other places were even more difficult to match, like Jembatan Jrebeng and Jembatan Legundi. Apparently these were different names for the same bridge, but it took a lot of time to figure this out. Validating and harmonizing these locations was an intensive manual task that required careful comparison of coordinates and names in google maps. Locations believed to be identical were set as such. From this, figure 2.1 was drawn up and numbered accordingly. A detailed list of locations is presented in annex A together with a larger version of figure 2.1.

2.1.3 Parameter Selection

Across all three agencies, 43 distinct parameters were recognized. For use in further analysis, it was decided that a parameter should meet at least the following requirements:

- They were present amongst all three agencies
- They had a significant amount of non-zero data and a low amount of zero data

The first is necessary because this research is interested in the 'hidden' differences between agencies and not in the obvious differences such as parameters. The second is necessary, because this creates mathematical problems during analysis. It is also not a very useful exercise to compare a bunch of empty values.

From the first requirement, only 17 parameters out of 43 remained. These are: TSS, DO, BOD, COD, TP, Fecal Coli, Total Coli, Temp, pH, EC, TDS, NO_3, NH_3, Deterg, Cu, NO_2 and Phenol. Based on the second requirement, the parameters of Copper (3520 null values), TDS (2801 null values), Phenol (2188 null values), Total Coli (1861 null values), Fecal Coli (1852 null values), and Detergents (1713 null values) were dropped. The other 11 parameters all had more than 80% of their entries filled with a numeric value and less than 1000 null values. As this was quite a large gap between these parameters and those other ones, this is where the threshold was placed.

The parameters that have thus been selected are:

- Temperature (Temp)
- pH
- Electrical Conductivity (EC)
- Dissolved Oxygen (DO)
- Total Suspended Solids (TSS)
- Biochemical Oxygen Demand (BOD)
- Chemical Oxygen Demand (COD)
- Ammonia (NH_3)

- Nitrite (NO_2)
- Nitrate (NO_3)
- Total Phosphorus (TP)

What these parameters represent in terms of water quality and how they are typically measured is elaborated upon in the next section.

2.2 SELECTED PARAMETERS

2.2.1 Temperature

Temperature is a physical parameter. It is also important for chemical processes in water as it speeds up the reaction speed and shifts certain chemical equilibrium's. It is measured in situ with a thermometer and can also easily be measured continuously.

2.2.2 pH

The pH value of a water body refers to how acidic/basic a water body is (USGS, Water Science School, 2018). It is a logarithmic scale that refers to the concentration of H^+ or OH^- particles in water. The scale goes from 0 to 14 where zero represents a concentration of 1 molar H^+ and 14 represents a concentration of 1 molar OH^-. Every step of 1 pH more means that the acidic concentration drops by a tenfold and the basic concentration is increased a tenfold. At a pH of 7 the two are balanced one to one. pH is very relevant to organisms as very small fluctuations can already make a large impact on their ability to live in it. Changes in pH can also be an indicator of certain pollutants. To measure pH, one can measure grab samples by means of titration in a lab, but more commonly, electronic sensors are used to measure in situ. These can suffer from inaccuracies as the calculations used to measure pH rely on certain conditions of other water parameters.

2.2.3 Electrical Conductivity

Electrical conductivity refers to how well water can conduct electricity. Pure water without other ions is a very good insulator as it is poor at conducting electricity (USGS, Water Science School, 2018). The electrical conductivity rises when ions are dissolved in it. As such, EC can also be used as a proxy for the total amount of dissolved solids as a linear relation between the two typically exists ((Atekwana et al., 2004)). Electrical conductivity is still a physical parameter of water quality and is usually measured in situ. It is also rather easy to measure continuously.

2.2.4 Dissolved Oxygen

Dissolved oxygen is a chemical parameter of water quality and refers to the amount of oxygen molecules that is dissolved in water. Oxygen in water is very important for aquatic live that consumes it (USGS, Water Science School, 2018). The capacity of water to hold oxygen is dependent on the water temperature where more oxygen can be dissolved in colder water. Dissolved oxygen can be measured with electronic sensors as well as lab equipment.

2.2.5 Total suspended solids

Total Suspendid Solids, or TSS for short, is an umbrella term for all particles in water larger than 2 µm (United States Environmental Protection Agency, 2012c). Anything that passes through a filter with pores of that size is considered a dissolved solid. Most common types of suspended solids are sand, silt and clay particles, plankton, algae and organic debris. Other types of non soluble solids are also included. TSS is commonly measured with grab samples where the sample is filtered and weighed. This method does have its limitations and more sophisticated methods such as acoustic doppler sensors have been under development (Wood, 2014).

2.2.6 Biochemical Oxygen Demand

Biochemical Oxygen Demand is an indicator of the amount of organic matter in a water sample. BOD specifically refers to the amount of oxygen that can be consumed by aerobic bacteria and other microorganisms to decompose organic matter (USGS, Water Science School, 2018). BOD is the measure for oxygen demand that most closely represents aerobic waste treatment and aquatic ecosystems (Boyles, 1997), but it is also the most time consuming test, with a standard test taking five days to complete. This means you can test for historic BOD values only and BOD measuring is not a useful tool for real time monitoring and intervention. There is also the problem that in waters polluted by toxic substances, BOD testing is less accurate, because microorganisms can not function properly (Boyles, 1997).

2.2.7 Chemical Oxygen Demand

Like BOD, COD is used to test the organic pollution in a water sample (Boyles, 1997). This can and is commonly used for wastewater, but can be extended to determine organic pollution in natural bodies of water. According to Boyles (1997), it tests for all the carbon that can be chemically oxidized. This includes BOD, but extends to organic compounds that cannot be oxidized by micro-organisms. It correlates with BOD on waste with constant composition, even if waste concentration fluctuates. The main benefit of COD testing compared to BOD testing is that results are produced in several hours as opposed to the five day period of BOD. It does usually require the environment of a laboratory, so samples still need to be collected and delivered to a central point.

2.2.8 Ammonia

Ammonia is an important nutrient for plant growth, as it is an important source of nitrogen (California Waterboards, 2021). Too much nitrogen in the environment can however have negative impacts. In surface waters, this comes mostly in the form of excessive aquatic plant and algae growth. This can have the consequence of oxygen deprivation during composition (USGS, Water Science School, 2018). Ammonia in the environment can exist in two forms: the un-ionized NH3 or the ionized NH4+. The dominant form is dependent on the pH (California Waterboards, 2021). There are many different methods for testing ammonia. One can use test strips, but more accurate measurements typically rely on optical colorimetric testing with some form of reagent (Li et al., 2020). Because this still relies on grab samples, more direct methods such as electrochemical sensors have been developed, but those are usually less accurate (Li et al., 2020).

2.2.9 Nitrite

Nitrite naturally occurs in the environment where it is a component of the nitrogen cycle. It is formed through oxidization of ammonium by bacteria (Hatzenpichler, 2012). It is then converted into nitrate by other bacteria (Daims et al., 2001). Nitrite is typically not as abundant as nitrate, because that is the more stable oxidation state. Most methods to measure nitrite involve colorimetry, which can only be done on grab samples (APHA, 1992).

2.2.10 Nitrate

Nitrate is the form of nitrogen most preferred by plants and is often an important component of fertilizers. Much like ammonia though, overabundance of nitrates in the environment can have damaging ecological impacts (USGS, Water Science School, 2018). A typical method used is cadmium reduction and colorimetry, though this will lump nitrates and nitrites together and can only be done with grab samples (United States Environmental Protection Agency, 2012b). Nitrate probes are also available, but these are more expensive, can suffer from interference and require careful calibration (United States Environmental Protection Agency, 2012b).

2.2.11 Total Phosphorus

Together with nitrogen, phosphorus is a very important nutrient for plant life. If both are available in environmental waters, then these are at serious risk of eutrophication and associated environmental damage (United States Environmental Protection Agency, 2012a). In nature, elemental phosporus is rare and it mostly occurs as part of the phosphate molecule (PO_4^{3-}) or in organic molecules. Measuring phosphate can be tricky, because concentrations are very low compared to other elements such as nitrate. It is still important though as those low concentrations can have major impacts (United States Environmental Protection Agency, 2012a). Total orthosphosphate refers to the phosphate ion in both suspended and dissolved form and it can be measured with the ascorbic acid method. A reagent is added to a grab sample and colorimetric methods are then used to determine concentration.

2.3 PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA) is a dimensionality reduction technique. What that means is that it can reduce information with a lot of different, but inter-dependable parameters into a few independent parameters (Dupont et al., 2020). By doing this, it also condenses information into a few important pieces. In this section, the methodology of PCA is first explained as well as some important assumptions about the underlying data. Because of those, scaling will be required and that will be explained in the next subsection as well as how PCA is applied to the dataset.

2.3.1 PCA Core Principles

The math behind PCA basically comes down to a change of basis for a given dataset. Provided a dataset with m variables, PCA will create m new variables, each a linear combination of the original variables. The weights to transform the original variables to the new variables are provided by the eigenvectors of the covariance matrix of the original dataset.

If there is a dataset X, with dimensions $m \times n$, where $n >> m$, then there is a covariance matrix of X, defined as $C = \frac{1}{n} X X^T$ (Shlens, 2014). The Principal components are the eigenvectors e of the determinant equation $|C - \lambda I| = 0$. The eigenvectors are then found by solving $C * e_i = \lambda * e_i$ for all $i = 1, 2, ..., m$ with the additional constraint that all eigenvectors form an orthonormal basis of size $m \times m$. As eigenvalues are naturally orthogonal, the only extra constraint is that for all eigenvectors: $||e_i|| = 1$ (Pérez-Bendito and Rubio, 1999). Once found, these eigenvectors e represent the loadings for the principal components. The principal components are sorted according to the size of the corresponding eigenvalue λ. The first principal component, or PC1, is thus given by the eigenvector e_i with $i = argmax(\lambda_i)$. The relative size of λ_i compared to the sum of eigenvalues equals the explained variance of the corresponding principal component.

The PCA is subject to the following assumptions about the dataset X (Shlens, 2014):

1. Linearity
2. Orthogonality
3. Large variances have important structure

The first assumes that the dataset has linear relations between parameters. PCA is a change of basis for the dataset and the new axis will line up with the greatest variances. These are linear axis though, so they will not be able to catch non-linear processes quite that well.

Orthogonality assumes that the relations in the dataset are orthogonal to each other. This becomes a problem when trying to establish multiple relations, because there is no guarantee that these do indeed behave orthogonally.

The last assumption basically states that principal components with a large associated explained variance also represent interesting relations in the dataset. This becomes problematic when a dataset contains a lot of noise, relative to the actual relations. In this case, PCA will fit for noise rather than for those actual relations. The relations found may then not represent actual relations within a dataset.

2.3.2 Data Preparation and Scaling

For data preparation, two things need to be done. First, the data needs to be scaled to equalize the importance of parameters. Second, the empty values in the dataset need to be filled, because the empty values would otherwise propagate into the principal components.

Data scaling is necessary for PCA, because otherwise parameters with large values dominate. It is also necessary to scale outliers, because otherwise distortion from a few outlier datapoints is very large.

As to what is the best method for scaling, a few were considered. First of all, the quite often used method of subtracting the mean and dividing by the standard deviation. This one would not be adequate though, because it does not consider the influence of outliers in this dataset well enough. Outliers would then need to be dropped and that would mean that information is lost as the outliers are not just statistical anomalies, but relevant information. Another method would be to scale between minimum and maximum, but this would still suffer the consequences of dramatic outliers and not properly adress that issue. From figure 3.4, it was found that the outliers required a different method of scaling compared to the bulk of the data, as even logarithmic scaling would not temper outliers sufficiently. Therefore it was decided that the best method would involve capping the data between a minimum and maximum value, setting values outside of this range to the minimum or maximum respectively and scaling the remaining data proportionally between those points. From 3.4 it was inferred that caps around the 1st and 99th percentile would work quite well to reduce the influence of outliers, while retaining as much information as possible. It was decided that data between those points should be scaled linearly, but a logarithmic version was tried as well. This, however, produced less satisfying results in PCA analysis, so linear was picked instead. More on this in section 4.4.3 of the discussion chapter.

All data is thus scaled linearly from 0 to 1 depending on the 1st percentile and 99th percentile of the data, which represent the values of 0 and 1 respectively. Doing so allows outliers to still contain value, but emphasizes on the bulk of datapoints and more general trends.

After scaling, the empty points are filled with the averages of the scaled parameters. By filling with the averages after scaling is applied, the empty points are filled with the origin of the new coordinate system that PCA creates. As such, these points will have no influence on the rotation of the coordinate system. Because the empty points need to be filled with the origin or the new reference system, it is important that filling happens after scaling. Because the mean after scaling is different from the mean before scaling, there would exist a vector going from the new mean to the scaled old mean. This vector would then influence the directions of the PCA. Because empty values carry no information, their influence on analysis should be minimized. Hence, it is important to fill after scaling.

2.4 NEURAL NETWORK PATTERN RECOGNITION

As part of the module Deep Learning for Hydrology and Water Resources Engineering of the CIE5431 Research Skills 1 course by Dr. Riccardo Taormina, a final exercise was submitted in which I created a neural network model that would classify the thesis dataset by agency. The results of this were so interesting that it was worth including it in the final thesis.

The basic principle of this exercise is to train an artificial neural network to recognize datapoints by agency. This means that the neural network is shown a datapoint from one of the three agencies discussed before and, on the basis of just the values for the eleven parameters, has to predict which of these three agencies took the measurement. No additional knowledge such as location or measurement time is provided. Based on these eleven values alone, the datapoint is thus classified as belonging to one of the agencies.

This exercise is a multi-class classification problem. Multi-class pattern classification is a tool of neural networks that can be used for a lot of different applications such as text document classification, image object recognition or handwriting digitization. Standard classification problems are two-class classification problems, i.e. either one class or the other. Multi-class classification is a fundamentally more complex technique, even for just three classes (Ou and Murphey, 2007).

In the next sections, the dataset will first be prepared for use in neural networks. Then the design process is outlined. In the final subsection, a reference model using logistic regression is introduced. This reference model will be used to indicate the utility of neural networks for this classification problem.

2.4.1 Preparing the dataset

To teach a neural network to classify datapoints successfully, the dataset must first be properly prepared. Preparing the dataset requires three things need to be done. First, the missing values need to be filled. Second, the data is split into three subsets and lastly, the data is scaled.

First of all, the missing values are filled. Because many components have a natural lower bound at zero and a nearly limitless upper bound for positive outliers, the median is chosen as the filling value, because for many parameters it better represents a 'standard' value than the mean.

Secondly, the total dataset is split into three subsets: a training set, a validation set and a testing set. This is done with a 60-20-20 ratio, i.e. 60% of data is used for training the neural network, 20% is used for validating the model and which iteration works best and a final 20% of data is used to test performance of the dataset. The datapoints are subdivided into each class randomly, but approximately these proportions of each agencies datapoints are present in each subset. The specific number of datapoints in each subset per agency can be seen in figure 2.2.

Finally, in order to better train the dataset, some scaling is necessary. Overall large values in one parameter compared to another will tend to have the neural nodes focus more on that parameter than on others, but for this exercise it is most effective if all parameters are considered equally. It is also necessary to demean the dataset, so the training algorithm works smoothly (Taormina (2022), personal communication). Unlike PCA, outliers are much less influential here, so a more simple scaling technique can be applied. All data is scaled by first subtracting the mean and then dividing it by the standard deviation, both of the training subset. This is applied to all subsets.

```
Training dataset info
-------------------------------------
x_train.shape: (3144, 11)
y_train.shape: (3144,)
There are 0678 instances of digit 0
There are 0849 instances of digit 1
There are 1617 instances of digit 2

Validation dataset info
-------------------------------------
x_val.shape: (1048, 11)
y_val.shape: (1048,)
There are 225 instances of digit 0
There are 298 instances of digit 1
There are 525 instances of digit 2

Testing dataset info
-------------------------------------
x_test.shape: (1048, 11)
y_test.shape: (1048,)
There are 248 instances of digit 0
There are 283 instances of digit 1
There are 517 instances of digit 2
```

Figure 2.2: Preparation of the dataset for use in the multi-class classification problem. It is split randomly in three parts: training (60%), validation (20%) and test (20%). Digit 0 matches with the EPA, digit 1 matches with BBWS and digit 2 matches with PJT.

2.4.2 Designing a Neural Network Model

After dataset preparation, the next task is to construct the artificial neural network. This will have to be done in several steps. The first is to decide on what type of neural network should be used. For this instance, it was determined that a feed-forward network would probably work best. Timestamp data is forgotten about for this exercise, so including time series is not an option and this is not image data, so there is no reasonable argument to include convolutional layers either. The final model will be a feed-forward neural network with dense layers.

The next decision is the type of activation function that will be used on each of the neurons of the dense layers. For feed-forward neural networks it is generally recommended to use the Rectified Linear Unit, or ReLU, activation function (Goodfellow et al., 2016, 6.1). It is worth investigating though if this is also correct for this model. Therefore, an initial model was created five times, each with a different activation function on the hidden layer. A hidden layer refers to one or multiple layers that are placed between the input and the output layer. The hidden layer consists of multiple neurons. In a dense layer, each neuron will present a new value based on a matrix transformation of all the values in the previous layer. For the first dense layer this matrix transformation looks as follows: $\mathbf{h} = actfun(\mathbf{W}^T\mathbf{x} + \mathbf{b})$, in which $\mathbf{x}$ represents the input vector and $\mathbf{h}$ represents the hidden layer. Subsequent hidden layers will just use the previous hidden layer as input instead. The weights are represented by the weights matrix $\mathbf{W}$ and bias vector $\mathbf{b}$. Those are what will be trained for. The activation function needs to be decided upon as well as the number and size of the hidden layers.

To find the most optimal activation function, these are thus tested with a single dense layer consisting of 100 neurons. In figure 2.3 the results of training a model with these different activation functions is shown. The output layer uses a softmax function, because that is best at handling a probability functions with more than 2 different outcomes (Goodfellow et al., 2016,

6.2.2). The softmax function that is tried in figure 2.3, refers to the activation function on the hidden layer, not the output.

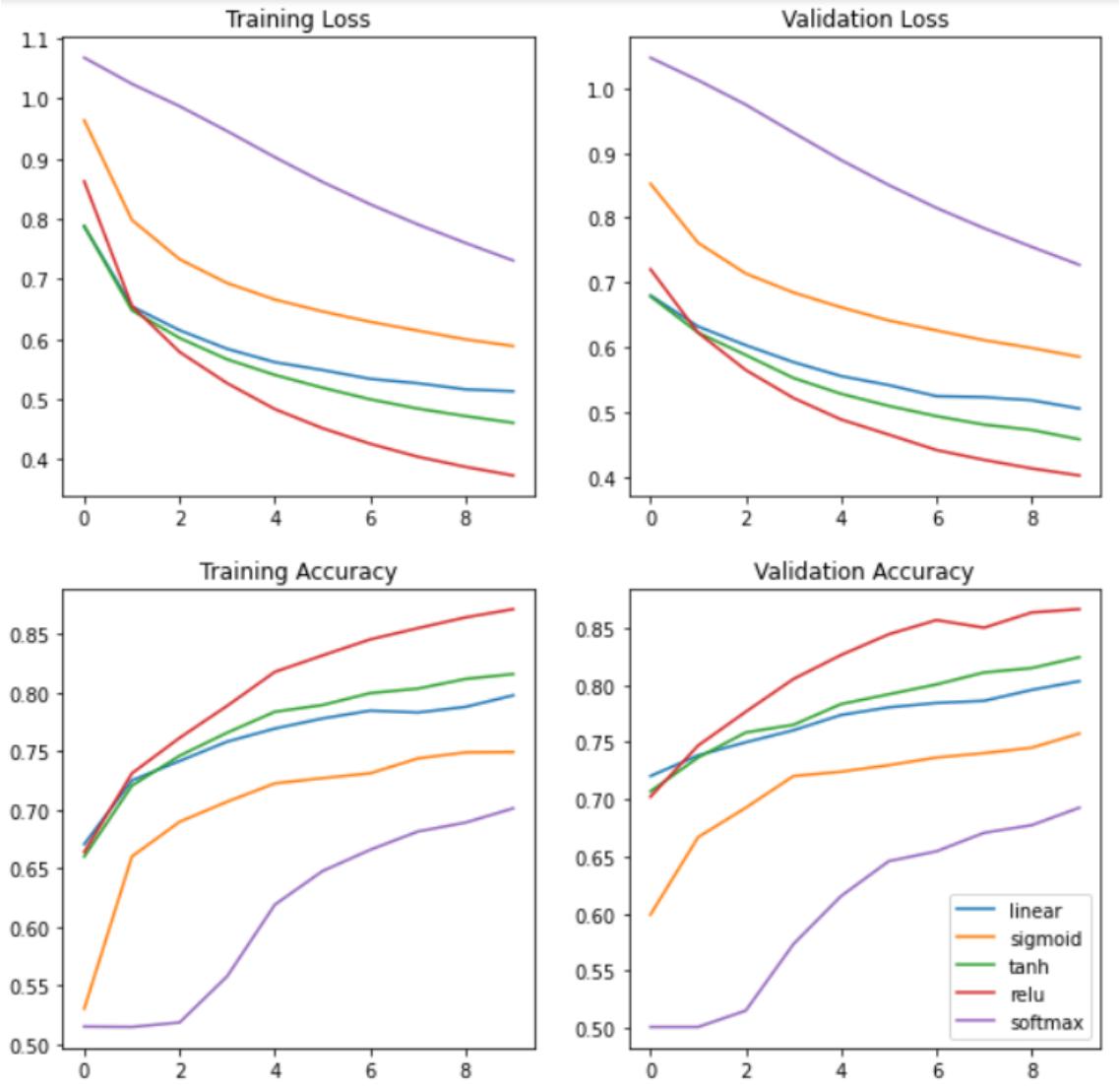

Figure 2.3: Comparison of different activation functions.

From figure 2.3 it can be seen that the ReLU function clearly performs best with the lowest combined loss during training and validation as well as the best accuracy. The assumption that this activation function would be the most optimal is thus correct.

Next, an optimal model architecture needs to be chosen. It was decided that not more than 100 total neurons should be used, so the question mostly came down to how to distribute those in hidden layers. For this, 4 alternatives were tried. One with all neurons in a single layer, one with neurons spread over two layers, one with 3 layers and one with 4 layers. For two layers, neurons are split up equally with 50 neurons in both. For three layers, the first two layers

contain 40 neurons and the third contains 20. For four layers, the neurons are divided in a 40, 30, 20, 10 order with 40 in the first and 10 in the last layer. The number of neurons per layer is descending, the reasoning being that the network aggregates data on every layer and would need less neurons on neurons to cover a lot of aspects for every level it passes through (Taormina, personal communication).

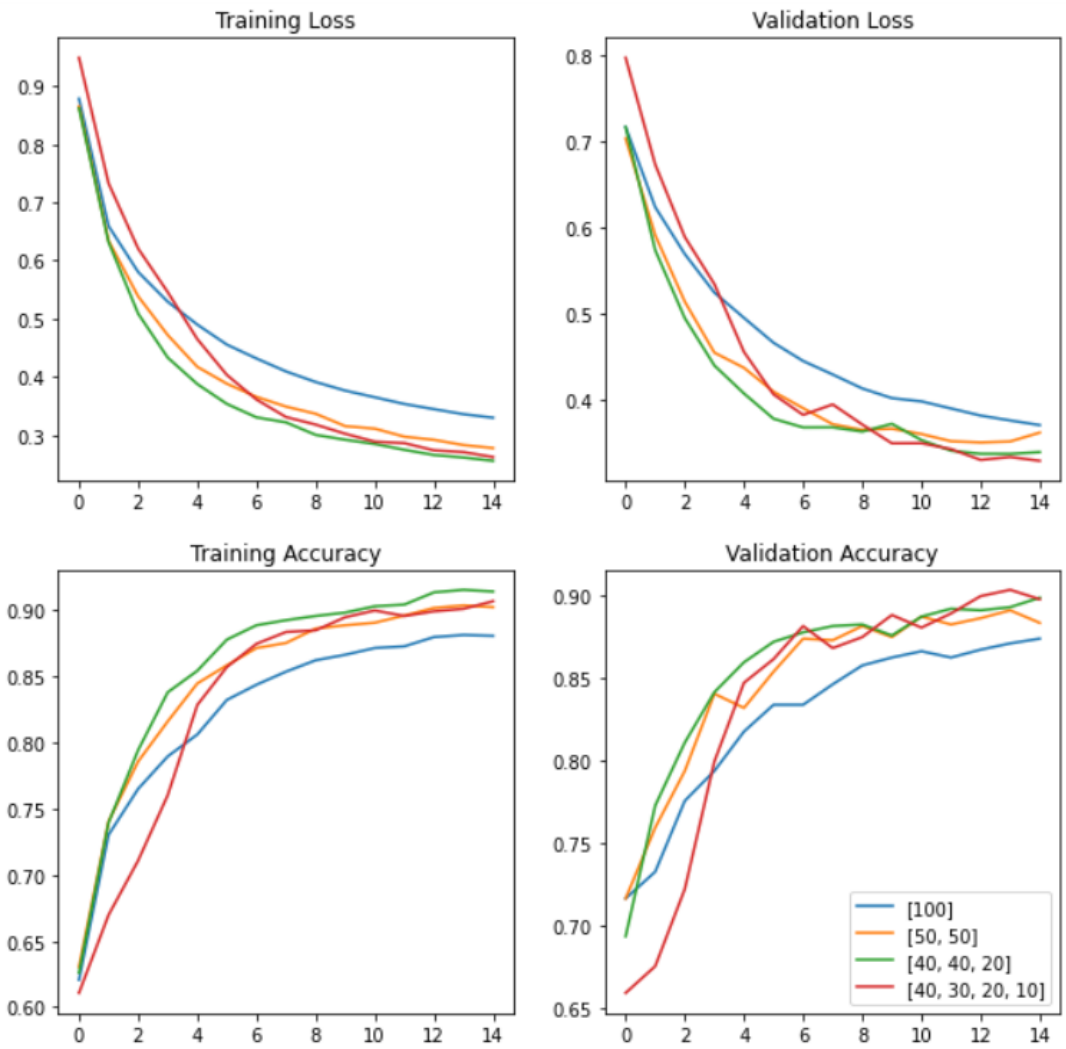

Figure 2.4: Comparing different model structures, with 1, 2 or 3 hidden layers respectively.

The results of the different layer architecture are showed in figure 2.4. The single layer clearly performs the worst. The model with two layers also does slightly worse. The models with 3 and 4 layers have very comparable results, but the model with 3 layers has more stable performance across different iteration stations. Therefore it was decided to continue with the [40,40, 20] structure.

```
Model: "model_10"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
Input (InputLayer)           [(None, 11)]              0
_________________________________________________________________
dense_19 (Dense)             (None, 40)                480
_________________________________________________________________
dense_20 (Dense)             (None, 40)                1640
_________________________________________________________________
dense_21 (Dense)             (None, 20)                820
_________________________________________________________________
Output (Dense)               (None, 3)                 63
=================================================================
Total params: 3,003
Trainable params: 3,003
Non-trainable params: 0
```

Figure 2.5: Overview of the selected ANN model for the classification problem. Param refers to the number of parameters that can be fitted to train the model.

An overview of this model with [40, 40, 20] structure is provided in figure 2.5. It consists of an input layer, the eleven water quality parameters, then two dense layers with 40 neurons and a dense layer with 20 neurons and finally an output layer that uses the softmax function. This output layer creates relative probabilities for the 3 different agencies. In the final step, the author with the highest probability is selected as outcome.

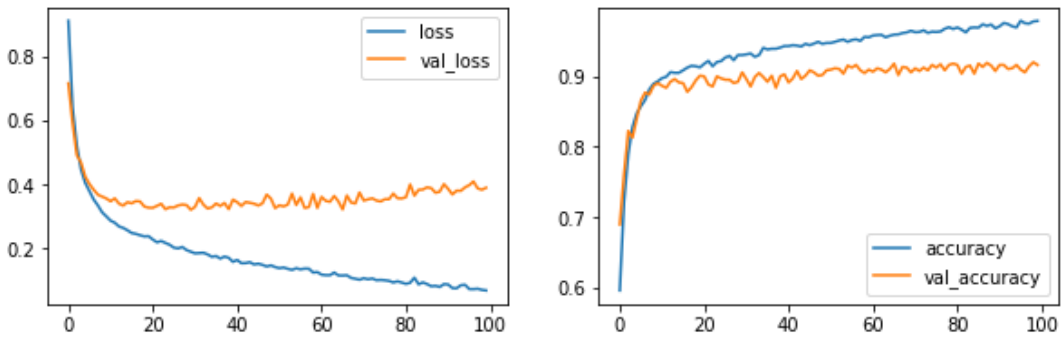

Figure 2.6: Model Performance during training and validation. Training performance is indicated with the blue line and validation performance is indicated with the orange line.

The model is trained using the Adam algorithm (Goodfellow et al., 2016, 8.5.3) for 100 epochs. The accuracy and loss function of training and validation is shown in figure 2.6. For training, the model was also given a callback and an early stopping function. Both monitor the validation accuracy during the training progress. Callback saves the model with the highest validation

accuracy and early stopping will stop training if no improved model is found for 50 consecutive epochs. After training has stopped, callback will then present the model with the best validation accuracy. Because this is not a completely unbiased selection of a model, to truly test its performance, the model will need to see another 'fresh' set of datapoints. This is why in the first step, the data was separated in three different sets.

Another thing to notice in figure 2.6 is that after some 10 steps, the performance indicators for training and validation start to diverge, so while the training keeps improving, the validation does not keep up with this improvement. This is symptomatic of overfitting and yet another reason to use early stopping and callback algorithms. These will select on the basis of the best validation accuracy of the last 50 epochs, however they do not necessarily prevent from selecting an overfitted model. Because of this small bias during selection, it is extra important to only interpret results from the models' performance on the test dataset, because that is the only set of data that did not influence the selection of the best performing model.

2.4.3 Logistic Regression

In order to establish the additional value of a deep neural network model, another reference model was created. This logistic regression model casts the input layer directly into the output layer without any hidden layers in between. The basic layout is presented in figure 2.7.

```
Model: "model_25"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 Input (InputLayer)          [(None, 11)]              0

 Output (Dense)              (None, 3)                 36

=================================================================
Total params: 36
Trainable params: 36
Non-trainable params: 0
```

Figure 2.7: Model Performance during training and validation

The number of weights that need to be trained (called Param in the figure) is 36. This is significantly less than the previous ANN model that used 3003 weights. The only mathematical operations that happen here is a matrix transformation followed by the application of the softmax function. The matrix transformation looks like this: $\mathbf{z} = \mathbf{W}^T\mathbf{x} + \mathbf{b}$. In this, $\mathbf{x}$ is our original datapoint with eleven parameters. $\mathbf{W}$ and $\mathbf{b}$ represent a 11 x 3 matrix and a vector of size 3 that comprise of weights found during training. The vector $\mathbf{z}$ is a size 3 vector that serve as input for the softmax function that looks like this: $softmax(z)_i = \frac{exp(z_i)}{\sum_j exp(z_j)}$. This is basically the same as the final step in the neural network where the last hidden layer is cast onto the output layer, but instead it casts the input straight into the output. The result here is that the output values

z_i that go into the softmax function are a direct linear function of the input parameters. This is also why this approach is chosen for comparison, because it somewhat represents what a human would be capable of given the input data. The basic assumption here is that logistic regression should thus be able to capture most of the obvious relations between parameter values and the agency they belong to. When in the final step, these results are compared to the neural network, the difference represents the non-obvious relations between datapoint values and the datapoint author.

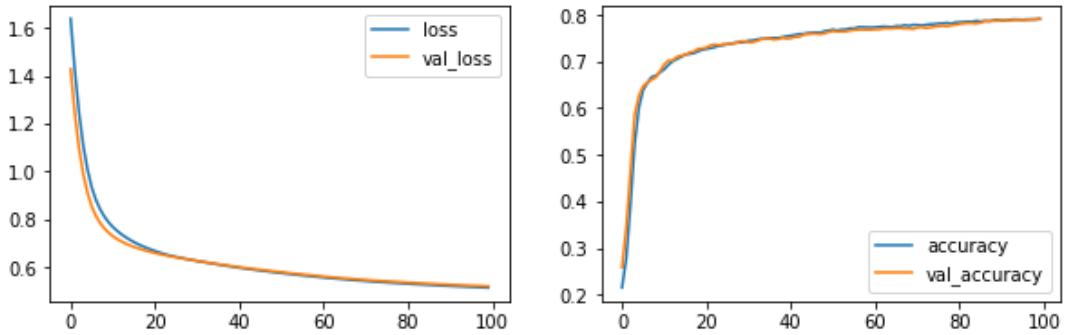

Figure 2.8: Model Performance during training and validation

The training of the logistic regression is done in a similar manner as the neural network and the performance of training is shown in figure 2.8. The training and validation results are very similar, indicating that the model is not overfitting and improvement of loss and accuracy decreases with each step. The same callback and early stopping algorithms are applied, though they will not really do anything this time, as validation accuracy still tends to improve with every epoch.

3 RESULTS

In this chapter results of analyses described in the methodology are provided with a short interpretation. These are subdivided in four main sections. Section 3.1 describes the water quality parameters of the Brantas in time and space. This is supported with boxplots of the data, which can also be found in Annex B. Section 3.2 goes more into the general statistics of the composite dataset. Principal component analysis is then presented in Section 3.3 and the chapter is closed off with the results of the neural network classification problem in Section 3.4.

3.1 DESCRIPTION OF GENERAL WATER QUALITY

Relations between measurement parameters in time, along the main stretch of the Brantas and in its branches have been plotted using box-and-whisker plots. This resulted in a total of 33 plots, three for every measurement parameter. They can be found in Annex B. Different boxplots are made for every authority along the different locations in the Brantas. For the boxplots in time, data of different authorities is aggregated together. A subplot is featured, illustrating the water levels at the Ploso bridge as measured by PJT, location 20 in the main trajectory plots respectively. In order to better detect more general patterns among these plots, some very large outliers have been left out of the y-axis window. In the next segment, the main relations are described.

As can be seen in Figure B.1, water temperature is initially low at the source, but rises rather quickly to over 25 °C near Malang and averages around 30 °C in the mid and lower sections of the Brantas. In the branches, as displayed in Figure B.2, no noticeable fluctuations exist, but both figures do show a pattern where the EPA seems to measure consistently lower temperatures than BBWS in the same places. The variations of temperatures over time, as displayed in Figure B.3, show slightly lower values during flood seasons and slightly higher values in the last quarter of the year, but temperatures are generally quite stable.

The pH tends to float between values of 6.0 and 8.5. As for the different locations, not much differences can be noted in Figures B.4 and B.5 other than that the pH starts off relatively low around the source and rises very quickly to values between 8 and 8.5 around the city of Batu. This then drops of and stabilizes between 7 and 8. More variation is found in time, as can be seen in Figure B.6 where pH usually drops during the rainy season and rises in the dry season. This is likely due to the fact that typical rainwater is slightly acidic. Air pollution could potentially make this effect even more pronounced.

Electrical conductivity is tightly connected to the amount of minerals dissolved in water and as far as the Brantas is concerned, variations in space seem to be not that large, as can be seen in Figure B.7, with two noticeable exceptions. The first are the low values around the source where

the water is still relatively unpolluted, resulting in a lower concentration of minerals and thus a lower EC value. The other big difference in this figure is the concentrations at location 20, Cheil Jedang ferry. This could be related to the high values for the various nitrogen components and phosphates found here as well (Figure B.22 and B.31). In Figure B.8, there is one location that measures exceptionally high values of electrical conductivity, that is location 60, Tlocor ferry. This measurement location is relatively close to the sea though, in the Porong branch of the Brantas, and seawater intrusion is likely responsible here. Variations over time in Figure B.9 are relatively modest and most likely correlated to the rainy season as well, with low points during this and peaks during the dry season.

Dissolved oxygen on the main Brantas stretch as portrayed in Figure 3.1 (B.10, large version) has some interesting relations in space. It is highlighted below, because oxygen is an important driver for many chemical processes in the river. Initially, the relatively clean water at the Brantas source contains high concentrations of oxygen, but those are considerably lower at the next point. Until the split between the Surabaya and Porong rivers, the oxygen content remains relatively stable with concentrations between 6 and 8 milligrams per liter with some low outliers. After the transition to the Surabaya river the situation gets worse, as low oxygen concentrations become more and more common and higher concentrations become rarer. This is a trend that is separately visible in all agencies' data and is thus an important process in this stretch of the river.

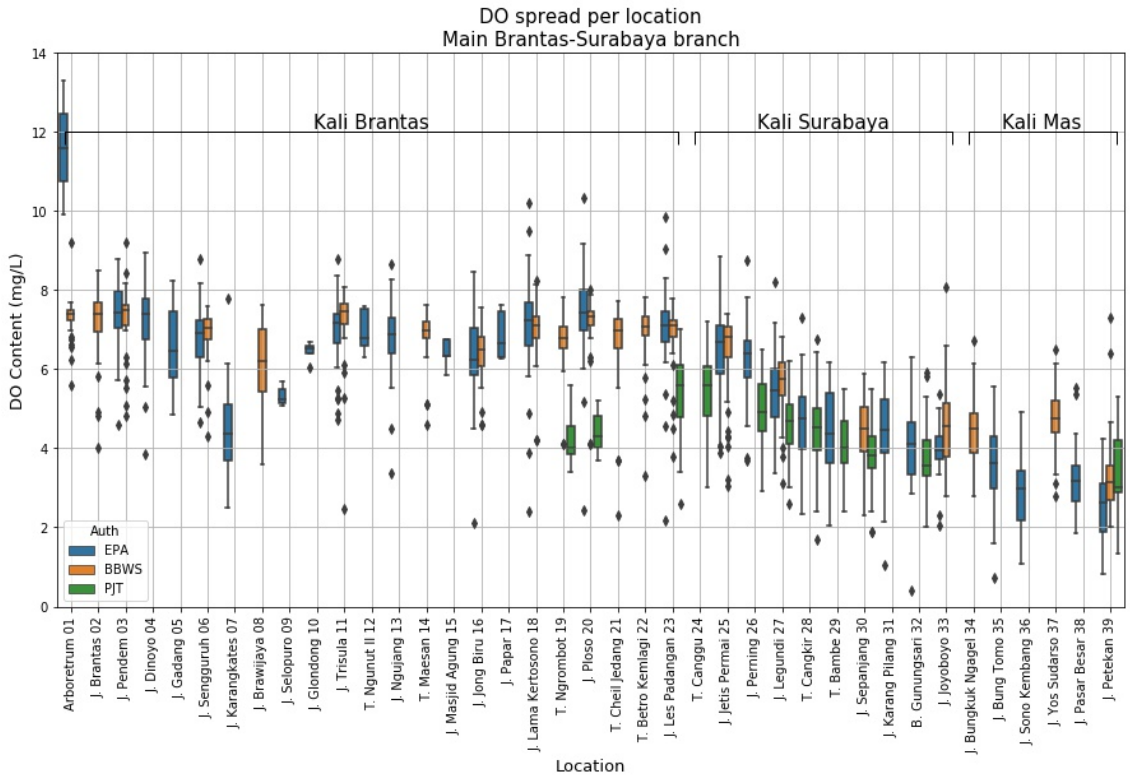

Figure 3.1: Dissolved oxygen concentrations along the main Brantas stretch.

The inflow branches of the Brantas and Surabaya rivers could in part be contributing to this drop in dissolved oxygen, as can be seen from figure 3.2 (enlarged in annex in figure B.11). Concentrations of dissolved oxygen are pretty reasonable at the earlier inflow branches, but much lower in the later ones. Especially the Kwangen, Tengah and Kedurus streams seem to often deliver oxygen deprived water to the main river. These are also the streams that are located in industrial areas which is a likely cause for this. Interestingly though, oxygen concentrations do not seem to be very dependent on time, as can be seen in Figure B.12. There seems to be no correlation to the river discharge variations in this sense.

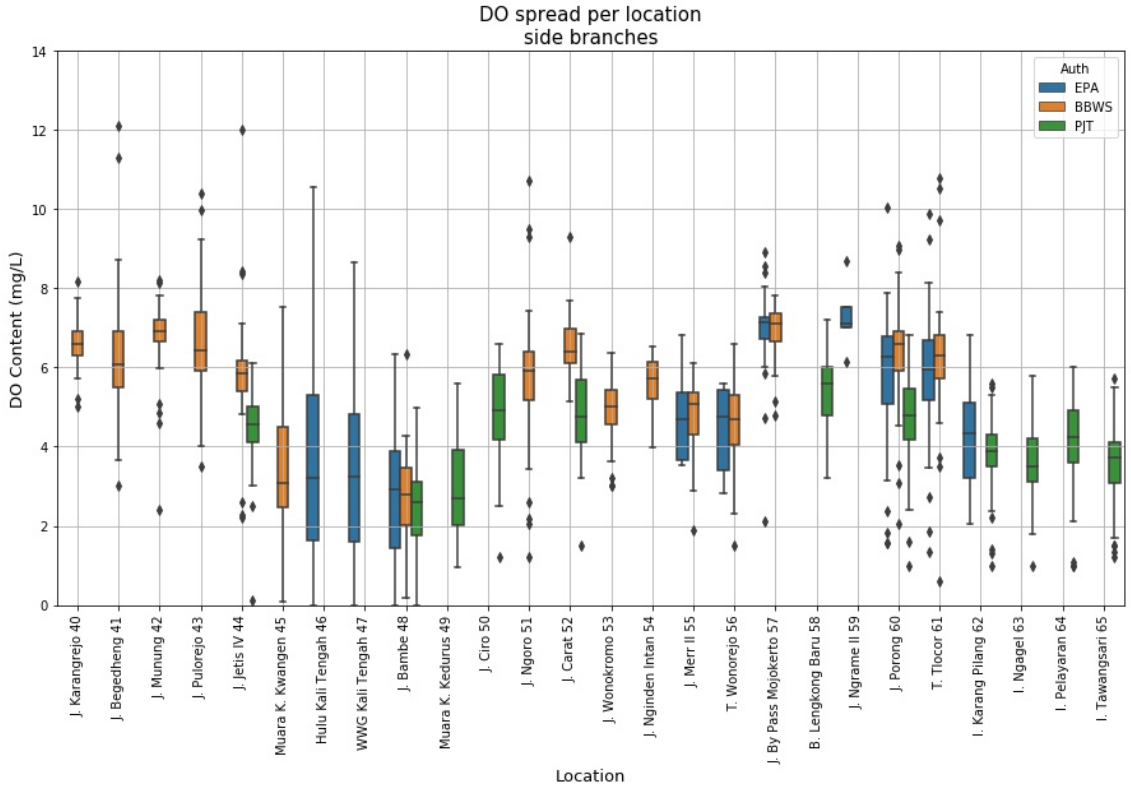

Figure 3.2: Dissolved oxygen concentrations in the branches of the Brantas.

Total suspended solids in Figures B.13 and B.14 seems to be fairly constant in space with it being slightly higher in the Surabaya river stretch and certain branches. Large drops are found downstream of large dams such as at location 7, Karangkates bridge. Most deviations of TSS are actually seen in time. In Figure B.15, which is also displayed below in Figure 3.3, suspended solids have relatively low concentrations throughout the year, but can rise spectacularly with increases in discharge during the monsoon season. This is expected as heavy rain storms will typically cause overland flow and flush all sorts of material and debris into the river streams. This effect is clearly visible in Figure 3.3

Variations in biochemical oxygen demand are seemingly not that dependent on location (Figure B.16) or time (Figure B.18). High and low values can seemingly appear anywhere at any time. That said, there are some differences when looking at some of the branches of the Brantas. In

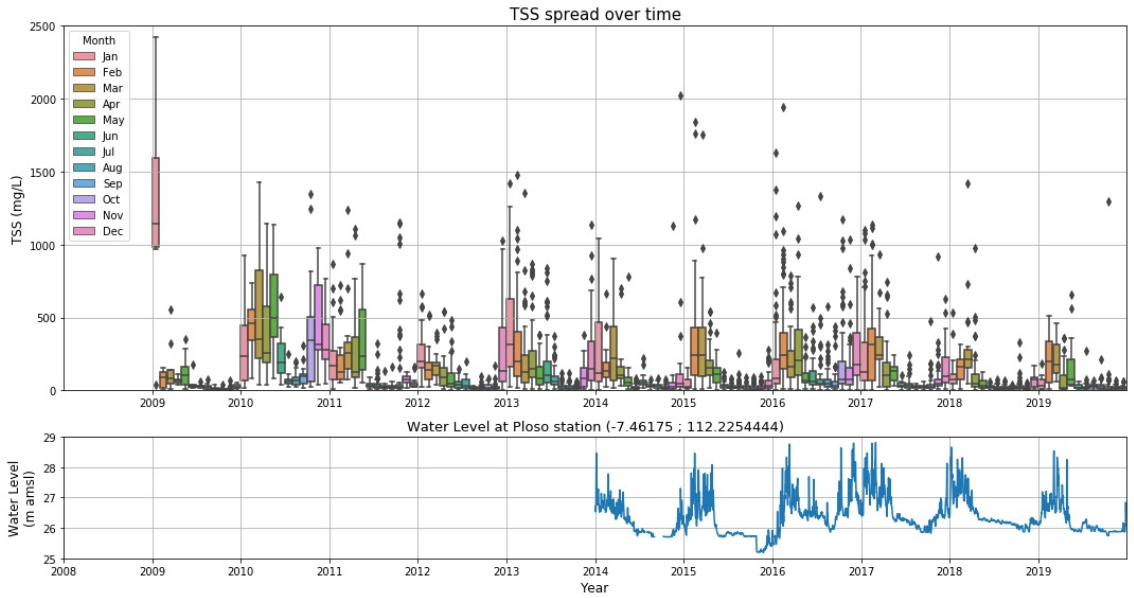

Figure 3.3: TSS concentrations in the Brantas over time.

Figure B.17, it is clearly visible that the Kwangen en Tengah streams, Bambe bridge is located at the outflow of the Tengah stream, regularly produce much higher concentrations than any of the other side branches.

Patterns in chemical oxygen demand are very similar to patterns in BOD. In the main stretch though there is a slightly more visible one. Figure B.19 displays a slight increase in COD concentrations in the lower parts of the Brantas and subsequent Surabaya and Mas rivers. Figure B.20 displays a similar pattern to Figure B.17 where the Kwangen and Tengah stretches again display many above average values. An interesting difference here is that at Tlocor ferry there is also a significant increase in COD that is not observed with BOD. This implies that the increased chemical oxygen demand is mostly not of a biological nature. In Figure B.21 the variations of COD in time are plotted and while there are some, they do not seem to be clearly linked to a seasonal component.

The spread of various nitrogen components is visible in Figures B.22 through B.30. These components are ammonia (NH_3), nitrite (NO_2) and nitrate (NO_3). Along the Brantas, these rise slightly along the trajectory with two notable exceptions. The first is at location 21, the Cheil Jedang ferry. Here concentrations of all nitrogen components are above average with ammonia concentrations being exceptionally high. The other is at Petekan bridge, the last point on the Mas river before it reaches the sea. What is interesting here is that the ammonia concentrations are rather high, but this does not seem to translate to nitrite and nitrate. A possible reason for this could be related to the dissolved oxygen concentrations, which as can be seen from Figure 3.1, are fairly low at this point. With a low supply of oxygen, ammonia could not be quite as easily converted to its higher oxidation states of nitrite and nitrate.

In the branches and sidestreams of the Brantas, nitrogen components follow a similar pattern to COD. Figure B.23 again shows high concentrations in the Kwangen and Tengah tributaries

as well as the Kambing tributary (Carat bridge, location 51). In Figure B.26, we see this trend continue for the Tengah and Kambing streams, but the Kwangen stream is surprisingly absent. Nitrate concentrations in the branches are shown in Figure B.29 and are a bit more constant. The Kwangen river however, still sees higher concentrations than elsewhere.

Figures B.24, B.27 and B.30 show the variations of these nitrogen components over time. The general trend is that concentrations are relatively stable throughout the year, peak at some point and then return to the stable level. For ammonia, this peak is usually reached in October and near the end of a dry spell. Nitrite shows peaks mostly in January prior to 2015, but in later years this is mostly November. Nitrate fluctuates a bit more year round, but with peaks falling anywhere between January and June.

The last parameter observed is total phosphates and the spread in space is fairly similar to our nitrogen components, as can be seen from Figure B.31, where no specific trends are seen except for the increased concentrations again at Cheil Jedang ferry. In the branches, Figure B.32 shows a similar pattern, where the highest concentrations are found at the outflow of the Kwangen and Tengah tributaries. As for the time dimension, Figure B.33 shows no clear peaks, but phosphate concentrations do seem to be lower during and at the end of the monsoon season. This indicates that the amount of phosphate discharged into the river is fairly constant, but this mass is diluted with more water.

3.2 DATA STATISTICS

Some of the core data statistics are shown in Table 3.1. Besides minimum, maximum, mean and standard deviation, a large array of parameter percentile values has been included as well. A wide range of percentiles has been delivered, because most parameters do not have a uniform distribution. For example, ammonia contains nothing but zeros in the first two percent. On the other hand, its maximum reported value is well over three times larger than its 99.9th percentile. This is not uncommon though among these measurements. TSS has a maximum reported value more than six times the 99.9th percentile value. This is not the case for all such parameters though. Electrical conductivity for example has a maximum reported value that is only about fifty percent larger than its 99.9th percentile but this 99.9th percentile is more than ten times larger than the 99th percentile. These distributions do not adhere to typical known statistical distributions that can be characterized with few parameters.

In addition to this table, a distribution plot is provided in Figure 3.4. Shown here is the distribution of the measurement parameters, but split by agency. With the exception of temperature, pH and dissolved oxygen, which are plotted on linear axes, the distribution plots are produced with a logarithmic x-axis. The 1st and 99th percentile have also been made visible by the black and red dotted lines. The red line for the 1st percentile is not visible in the graph concerning NO_2, because it lies at the value of zero, which cannot be visualized on a logarithmic scale.

As for the distributions themselves, like was shown in Table 3.1, most are not distributed according to a known statistical distribution. The shape of the cumulative distribution curve for a certain parameter is also not always consistent between agencies. Nitrate for example has all authorities on a fairly similar median, but in the lower percentiles, the EPA reports slightly higher values. In the higher percentiles, this is done by BBWS. These differences are still very subtle,

	Temp	pH	EC	DO	TSS	BOD	COD	NH_3	NO_2	NO_3	TP
Min	7.51	3.2	5.46	0	0	0	0.1	0	-0.001	0	0.003
0.1st p	16.9	6	23	0	0.9	0.02	1	0	0	0.009	0.005
1st p	22	6.35	162	0.96	3	1.2	2	0	0.002	0.158	0.021
2nd p	24.3	6.51	207	1.3	4	1.55	3	0	0.002	0.249	0.031
5th p	26.7	6.73	277	2.15	8	2	5	0.001	0.004	0.474	0.054
10th p	27.4	6.87	316	2.92	11	2.4	6.72	0.005	0.008	0.807	0.076
Q1	28.4	7.12	380	3.7	22	3.3	9.62	0.022	0.022	1.43	0.126
Median	29.2	7.39	461	4.8	53	4.65	15	0.099	0.066	2.05	0.196
Q3	30.1	7.68	550	6.2	162	6.92	24.2	0.247	0.15	2.6	0.44
90th p	31.1	7.9	760	7.21	366	12	37.2	0.56	0.282	3.3	1.04
95th p	31.9	8.02	1148	7.53	580	17	51.9	1.19	0.448	3.81	1.51
98th p	32.6	8.19	1501	8.04	850	28	88.6	2.8	0.715	5.3	7.07
99th p	33.1	8.33	1727	8.54	1036	37.8	140	4.99	0.901	6.79	12.6
99.9th p	35	8.67	23084	10.8	1917	185	669	19.1	4.09	19.3	32.9
Max	49.5	8.8	38500	17	12639	1182	2640	73	11.1	52	53.6
AVG	29.2	7.39	591	4.92	142	7.01	23.2	0.369	0.131	2.16	0.705
SD	1.94	0.415	1327	1.72	284	20.5	58.2	1.8	0.291	1.6	2.57

Table 3.1: Statistics of the composite dataset. Included are average, standard deviation, minimum, maximum and several percentiles.

but Figure 3.4 also shows the more pronounced differences between agencies. Phosphates values for example are shown to be distinctively higher for BBWS. The curve from BBWS here is also not in a neat s-shape. Ammonia is another notable graph as the entire distribution of the EPA is similar in shape to that of PJT and BBWS, but shifted to the left by a factor of almost one hundred. In the graph of ammonia, the curve of BBWS behaves in a stair like manner. This is due to the number of decimals that is reported. The difference between 0.01 and 0.02 looks very large on a logarithmic axis even though the true difference is relatively small.

There are many differences though between agencies. Only for temperature and electrical conductivity, the cumulative distribution functions are very close across agencies. BOD sees the EPA and PJT measure very similar values, but BBWS has an s-curve that is less steep. Other parameters show similar trends where the distribution curves will at some points coincide, but gradients and averages are slightly dissimilar.

Plotted in Figure 3.5 is the correlation matrix of the dataset. Two different matrices have been plotted, one where parameters have not been scaled and one where the same scaling algorithm as for PCA analysis has been used. This algorithm mitigates extreme values and will therefore emphasize relationships in the large mid-range scatter of data points, whereas the general correlation matrix will be swayed more by relationships between outliers. Because most outliers are not that correlated, most correlation coefficients are low in this diagram and relations are mostly found with parameters that do not have very extreme outliers such as dissolved oxygen or total phosphates. Because the one percent scaling algorithm is also applied in PCA, this correlation matrix is slightly more interesting. Most parameters still have a strong correlation to dissolved oxygen, but now electrical conductivity also shows some sort of correlation with a lot of differ-

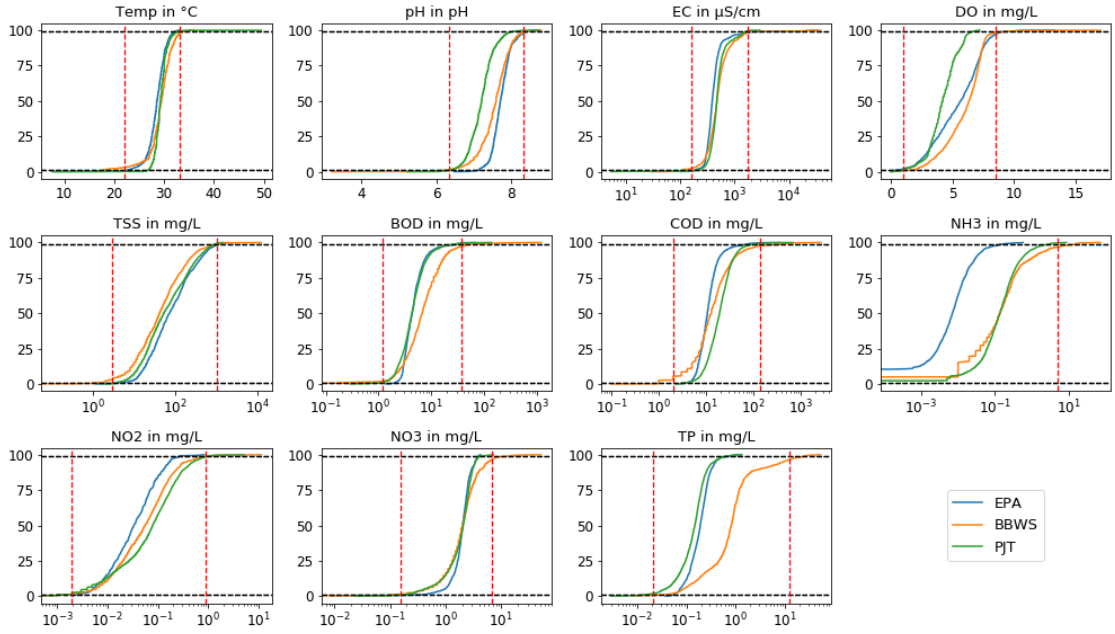

Figure 3.4: Distribution plot of measurement parameters. Parameters are split by agency. The black dotted lines indicate the 1st and 99th percentiles. The red dotted lines correspond to the values of these percentiles. Temperature and pH have been plotted on a linear x-axis, the other parameters are plotted on a logarithmic x-axis.

ent parameters, most notably with COD, BOD and NH3. Another strong correlation is the one between BOD and COD which was somewhat expected.

Some of these correlations have been outlined in Figure 3.6. A fit was made to estimate how well one parameter can predict the other. The parameters plotted on a logarithmic axis are fitted both by a linear relationship, as well as logarithmic one. This is according to the relationship: $\log \hat{y} = A \log x + B$. This was not done for the relationship between dissolved oxygen and temperature, because a linear axis made more sense here. The plotted relationships have been selected, because they showed some sort of interesting correlation, not just in the correlation matrix, but in the scatter cloud as well.

The first relationship in the top left corner of Figure 3.6 is that between BOD and COD. As both represent some sort of oxygen demand it is not strange that they are positively correlated. While they have a correlation coefficient of 0.61 without applying scaling algorithms, it turns out that one is not a good predictor for the other as the coefficient of determination is only 0.368 for the linear fit and 0.328 for the logarithmic fit. Despite these low determination coefficients, it is the strongest relationship between two measurement parameters in this dataset.

Two relationships that are also of interest, are the ones between electrical conductivity and BOD/-COD. These have been depicted in the top right and middle right graphs of Figure 3.6. The relationship between COD and EC is slightly stronger than that between BOD and EC. In the scatterplots of EC-BOD and EC-COD most points are located in a round, uncorrelated circle. Outside of this circle extends a panhandle of points to the right and up. These are the points

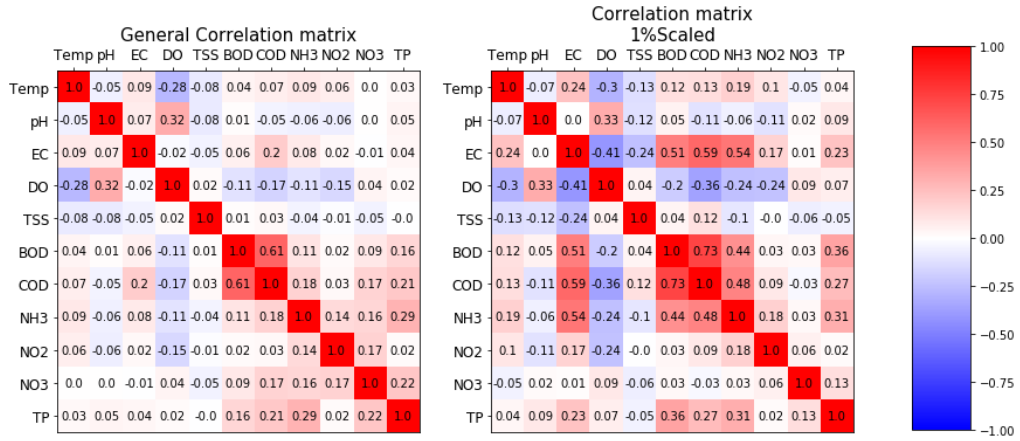

Figure 3.5: Correlation matrices of the dataset. On the left is the correlation matrix when no scaling is applied to the data. On the right is the correlation matrix where data has been scaled between the 1st and 99th percentile.

that drive the correlation. Apparently, for high values of EC and COD a relationship exists. This is similar between EC and BOD, but values of BOD are only slightly above average when EC values go above a certain threshold.

One cross plot is not set on a logarithmic axis. This is the one between temperature and dissolved oxygen, visible in the middle left graph of Figure 3.6. For most points in this cross plot, no clear relationship between the two exists, but a negative correlation is caused by a panhandle of low temperatures. It appears that low temperatures nearly always coincide with a high oxygen content around 7-8 mg/L. Between 27-33°C, oxygen concentrations are rather unpredictable.

Also depicted are the relations of EC to ammonia and total suspended solids, which are in the bottom left and bottom right graphs of Figure 3.6 respectively. Ammonia correlates positively with EC and suspended solids correlates negatively. They also have two things in common, the first being that the scatter consist of multiple somewhat correlated clouds. The second is that the logarithmic and linear fits are quite different from each other. This could either indicate poor true correlation or that there are multiple different correlations. As there seem to be multiple correlated clouds, this may indeed be the case.

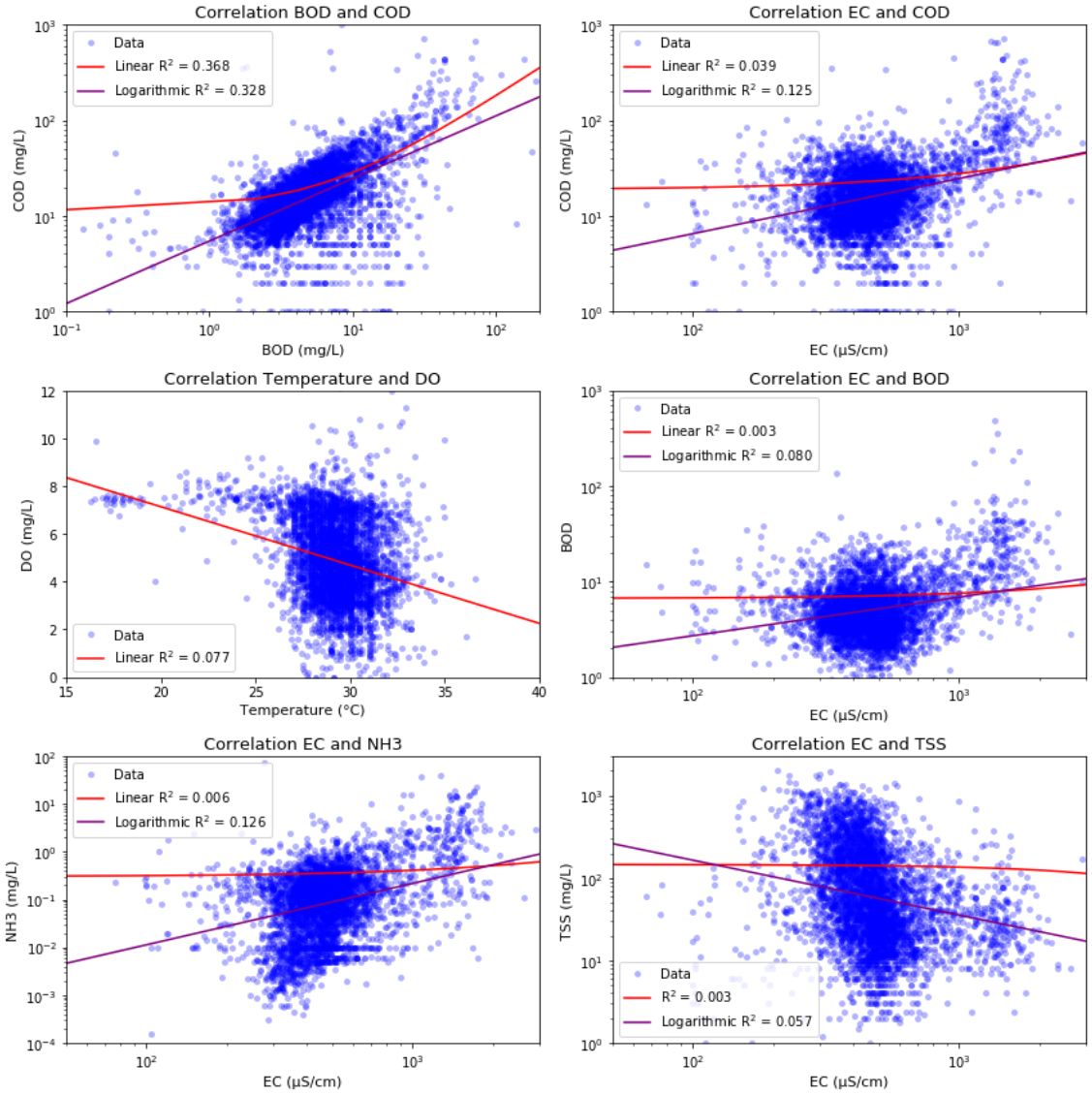

Figure 3.6: Cross-plot of the more interesting relationships. Except for the correlation between temperature and oxygen, the relationships are plotted on logarithmic axes. Besides plotting the parameters against each other, there has also been plotted a best fit. For those parameters on the logarithmic axes a logarithmic fit has been portrayed besides a linear fit. The R^2 has also been indicated.

3.3 PRINCIPAL COMPONENT ANALYSIS

3.3.1 PCA of composite dataset

The principal component analysis of the composite dataset has been visualized in Figure 3.7. The results consist of two important parts. First, the capture of explained variance and second, the loadings of principal components. The explained variance is a measure of how much of the variance in the parameters of a measurement point can be explained along a single axis. The principal components are ranked based on this capture of explained variance. As principal component analysis is a dimensionality reduction technique, we want to make as small of a selection as possible in PC's that together capture a significant amount of variance in the dataset.

To this end, there are two breaking points that are interesting, where involving another PC would contribute significantly less than the previous PC. This is after the third PC, as the first 3 PC's make up 56.2% of explained variance, or after the seventh PC as those seven make up 87.7% of total explained variance. The next principal component, PC8 would only contribute another 4% in explained variance.

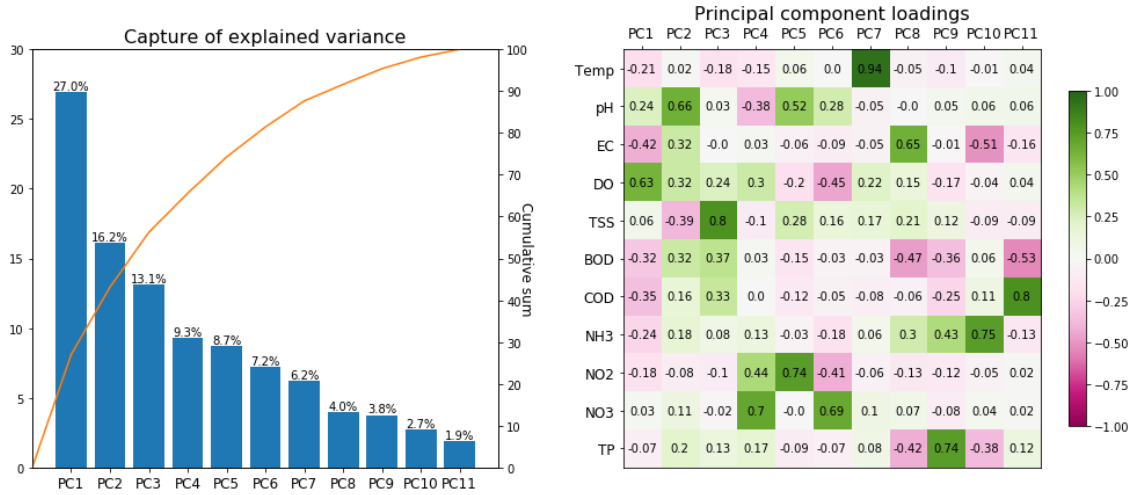

Figure 3.7: PCA Analysis, on the left is the capture of explained variance by principal component and on the right is the composition of said principal components, these are the loadings. Colors have been added to emphasize the larger absolute loadings. The largest loading in absolute terms has been defined as positive for every PC and is thus portrayed with a dark green color.

Looking at the PC loadings gives an idea of the important processes in the river. The first principal component is dominated by oxygen. It is inversely related to electrical conductivity, BOD and COD. The relationship between BOD/COD and oxygen is expected as BOD and COD represent the potential to consume oxygen. Higher values here will typically result in an increase of oxygen consumption which reduces the oxygen content. For the Brantas river, this relationship is the most powerful explainer for its water quality. What is interesting, is that electrical conductivity is also involved. While there is no direct chemical relation between EC and dissolved oxygen, there is a relation between EC and COD as soluble matter causes water to be more conductive.

BOD or COD can be such a form of soluble matter, though both can be of a non-conductive nature as well. That these are highly correlated in the PCA is in line with the correlation matrix of Figure 3.5, but not necessarily an expected outcome with regards to general knowledge about relations between EC and BOD/COD (Van Breukelen, 2022).

The second principle component is dominated by pH and is inversely related to total suspended solids. Figure 3.3 showed that TSS was highly variable in time and high values occurred almost exclusively during the monsoon season. This would explain why it is related to pH, as rain is naturally more acidic than river water, because acidic substances mostly originate from the atmosphere and alkaline contents originate more from sediments, rocks and certain industrial discharges (Partnership, 2022). A peak in discharge due to rainwater will thus cause the pH to drop and the solids to increase. Additionally, due to rainwater's relative purity, other components in the river may get diluted. This relationship is visible in PC2 in the loadings of BOD, COD, TP and EC. However, most rainwater in the river during the monsoon is overland flow. This rainwater is in contact with the top soil, which not only causes increases in TSS, but can increase the concentrations of these other components as well. This is likely why oxygen has a positive loading, implying its levels go down with rainfall. While rainwater itself has a high natural oxygen content, pollutants in the run-off can decrease this again. Clear relationships between the monsoon season and oxygen content are absent in Figure B.12 which shows the distribution of oxygen in time. This implies there are multiple mechanisms at work and hence it is hard to say anything conclusive about the relationship between PC2 and oxygen.

The third principal component is centered on suspended solids and additionally carries heavy weights in both BOD and COD. These were negatively related in PC2, but carry a positive relation in PC3. This indicates that we are dealing with multiple types of suspended solids or multiple mechanisms. PC2 is more focused on the natural flushing effect of rainwater where PC3 is more focused on the solid waste that ends up in the river. The natural flushing effect will happen everywhere in the delta, but the flushing of solid waste will only happen in those areas with a lot of waste on the streets.

PC4 through 6 is loaded mostly by pH, nitrite and nitrate. The loadings often counteract and actual correlation was also shown to be low in the correlation matrix. This implies that there may be a large amount of variability, but this is not necessarily related. Henceforth, PC4 through 6 does not provide more significant meaning than the parameters themselves.

PC7 is dominated primarily by temperature, but furthermore independent of other parameters. This indicates temperature as a parameter that is mostly independent and does not have many interactions with the water quality.

3.3.2 PCA by Authority

Besides a PCA for the composite dataset, PCA's were made for the individual authorities as well. The results of these are seen in Figure 3.8. The graph in the top left displays the capture of explained variance. The first principal component of the EPA captures 30.9%, that of BBWS captures 28.0% and of PJT captures 26.4%. The EPA has the most important first three components which together account for 60.2% of explained variance. This is 53.9% and 55.0% for BBWS and PJT respectively. The first six principal components seem to capture quite a large amount of

explained variance. This is 83.0% for both the EPA and PJT. BBWS scores slightly lower with only 78.7% accounted for by the first 6 principal components.

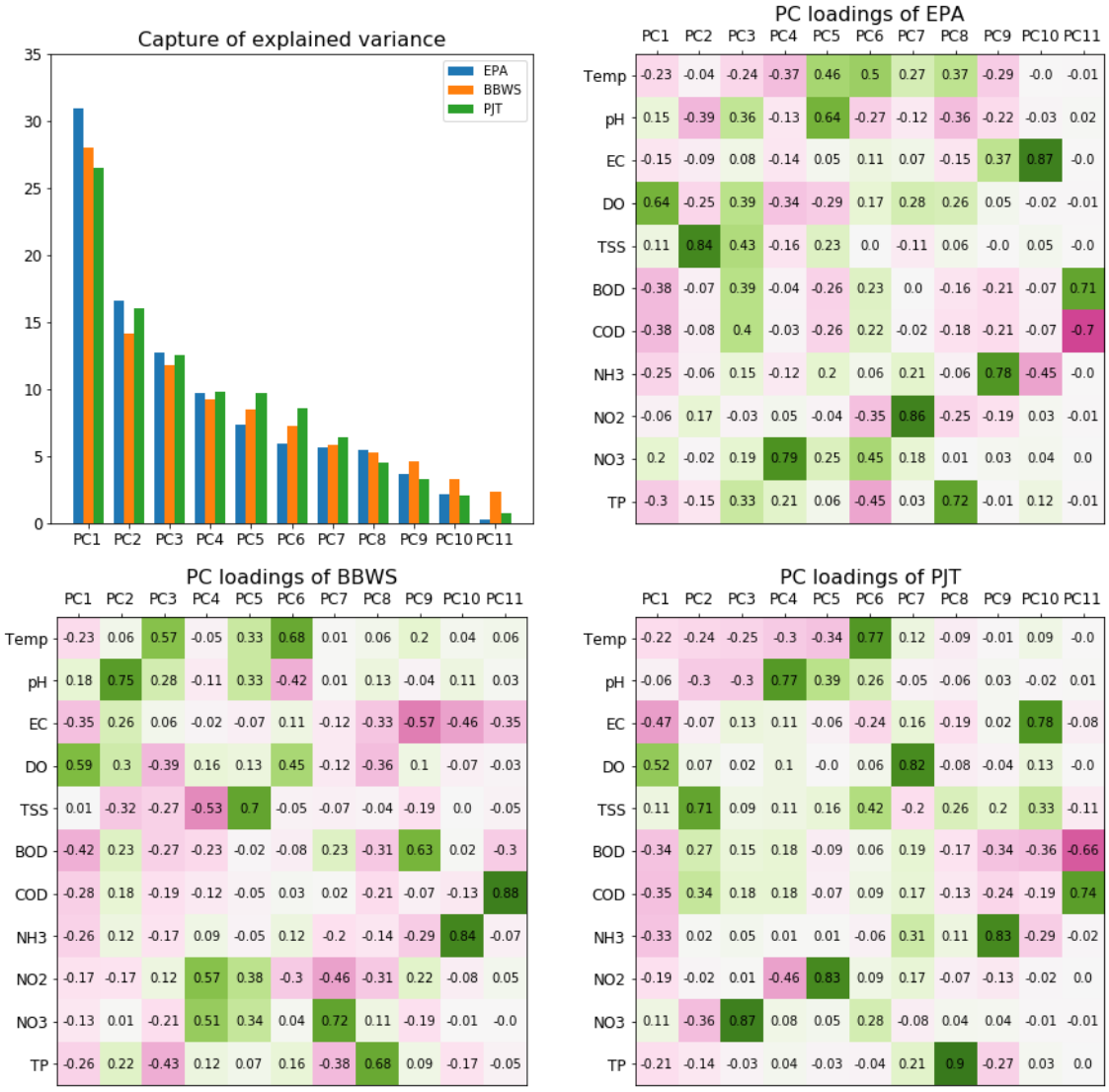

Figure 3.8: PCA of separate authorities. The top left graph depicts the capture of explained variance of the three different PCA's. The other 3 graphs depict the PCA loading of the specific authority.

The first principal component is relatively similar across the three agencies. Its heaviest loading is from dissolved oxygen which is inversely related to BOD, COD and NH3. It is also inversely related to EC, but the magnitude of this loading varies. PJT has a relatively large focus on EC while it is not very important to the EPA.

Larger differences appear when looking at the second principal component. The dominant loading for the EPA and PJT is total suspended solids, while it is pH for BBWS. There is also a remarkable difference between the EPA and PJT, as the EPA only sees a major secondary loading

from pH, while PJT has major loadings of COD and nitrate as well. Both still have an inverse relationship with pH, which indicates that this principal component is mainly driven by rainfall. BBWS's second principal component has its largest loading on pH. This is still inversely related to TSS and as such it describes the same rainfall process, but the emphasis has shifted.

The third principal component is where very large differences appear in the loadings. For PJT, PC3 is heavily related to nitrate and not much else. BBWS has its third component dominated by temperature, which is inversely related to phosphate and oxygen. For the relation between temperature and oxygen it acts as a magnifying glass to PC1, but as a counterbalance for the relation between phosphate and temperature. Therefore, attaching physical meaning to BBWS's PC3 becomes quite hard. PC3 is quite different though for the EPA. Here pH, oxygen, suspended solids, BOD, COD and total phosphate all have significant, but roughly equivalent loadings. This is likely to be the process of solid and biological waste being flushed in some areas of the river basin in addition to the natural monsoon processes. It is subordinate to the main processes of oxygen interaction and rainfall, but still important.

3.4 NEURAL NETWORK REPRODUCTION

The artificial neural network was tested with a fresh part of the dataset that was not used for training or validation. The confusion matrix of Figure 3.9 shows the performance of the model in this test dataset. In this confusion matrix, the true class is the actual author of the datapoint. The predicted class is the author this model thinks the datapoint is from. The datapoints that are predicted to be from the same author as the one that they are actually from, are found on the main diagonal. In Figure 3.9 most points are located on this diagonal. The largest group of errors occurred with datapoints from BBWS that were interpreted to be from PJT.

Based on this confusion matrix a set of performance indicators were calculated. These are shown in Table 3.2. The most significant performance for how well this model performed overall is accuracy, which was found to be 0.92. This means 92% of points was correctly identified in the right class.

	precision	recall	f1-score	support
EPA	0.90	0.95	0.93	248
BBWS	0.90	0.86	0.88	283
PJT	0.94	0.94	0.94	517
accuracy			0.92	1048
macro avg	0.91	0.92	0.91	1048
weighted avg	0.92	0.92	0.92	1048

Table 3.2: Classification report of ANN model with 40-40-20 structure.

Precision of a class is the fraction of points in a predicted class that are actually a part of that class. In a confusion matrix like figure 3.9, this would be number of points on the diagonal divided by the number of points in the column. For example, in figure 3.9, 518 datapoints are in the column belonging to PJT. These are all the points the model has predicted as a datapoint

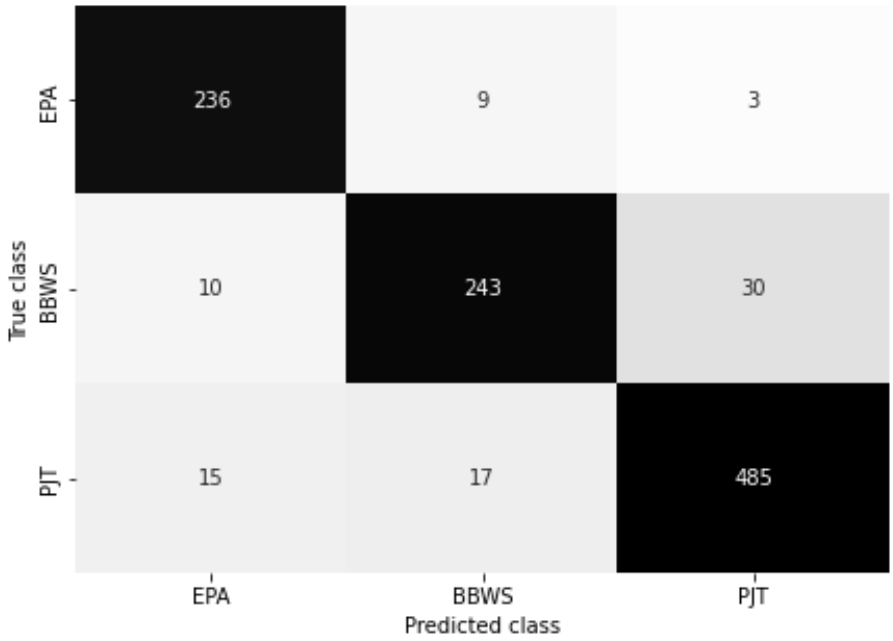

Figure 3.9: Confusion matrix of ANN model with 40-40-20 structure.

of PJT. From these 518 points, 485 were true datapoints from PJT. These are located on the main diagonal. The other 33 points in the PJT column were predicted as PJT, but were datapoints from BBWS or EPA. These were misclassifications. The precision of a class like PJT, can thus be seen as the reliabilty of the prediction PJT. In this case it is calculated as follows: $485/518 = 0.94$

Recall is the fraction of datapoints from a given class that are predicted to be in that class. In the confusion matrix of figure 3.9 this is represented as the number of points on the main diagonal divided by the number of points in the entire row. For example, the neural network was fed 248 datapoints from the EPA. From those 248, it classified 236 correctly as EPA and 12 as BBWS or PJT. The recall of a class like EPA, can thus be seen as how successful a model is at recognizing the points from the class EPA. For this case, the calculation is thus as follows: $236/248 = 0.95$

The F1-score is the harmonic mean of precision and recall and is calculated as: $F_1 = \frac{2*precision*recall}{precision+recall}$. In this scenario we can interpret it as a measure of how good the neural network is at recognizing the signature of a certain author.

There are two extra categories in Table 3.2, which are the macro average and the weighted average. Macro average is the overall average of this performance indicator, e.g. the average precision of all three classes. The weighted average is the average of this indicator weighted by the number of samples in this class.

All performance indicators are generally very close and range between 0.86 and 0.95. The performance indicators are overall slightly higher for the EPA and PJT and slightly lower for BBWS. As such it can be concluded that this model is better at recognizing the signatures of the EPA and PJT and slightly worse at recognizing that of BBWS.

Just because our model has an accuracy of 92% does not mean that the model is 92% successful at determining in which class a certain datapoint falls. We have to account for the fact that some of the datapoints are classified correctly based solely on luck. As there are three classes, any point has a 33% chance to accidentally be classified correctly. This model misclassifies 8% of datapoints. Assuming that for these are points our model could not detect any signature of and are therefore classified based on chance, then only 2 out of 3 of these datapoints will actually be misclassified. The 8% that is misclassified is thus likely part of a group of 12% that was classified purely based on chance. This leaves us with a group of 88% that the model could successfully pick up a signature from.

In order to evaluate the additional effectiveness of a complex neural network, the same classification problem was also posed to a logistic regression algorithm. The results of this are portrayed in a confusion matrix in Figure 3.10 and a classification report in Table 3.3. The confusion matrix clearly shows that this is significantly less effective as the cells are grayer and less black-and-white. The poorer performance is also visible in the classification report of Table 3.3. The total accuracy of the classification problem is just 0.76 or 76% for this logistic regression where it was 92% for the more complicated neural network. Recall and precision vary more for this classification model as recall for the EPA is only 0.54 while it is 0.90 for PJT. Precision is highest for BBWS at 0.85 and lowest for the EPA as well with 0.63. With an f1-score of only 0.58, this logistic regression model is thus pretty poor at recognizing the signature of the EPA. The individual scores for BBWS and PJT are better, but in no single performance indicator is the logistic regression competitive with the neural network.

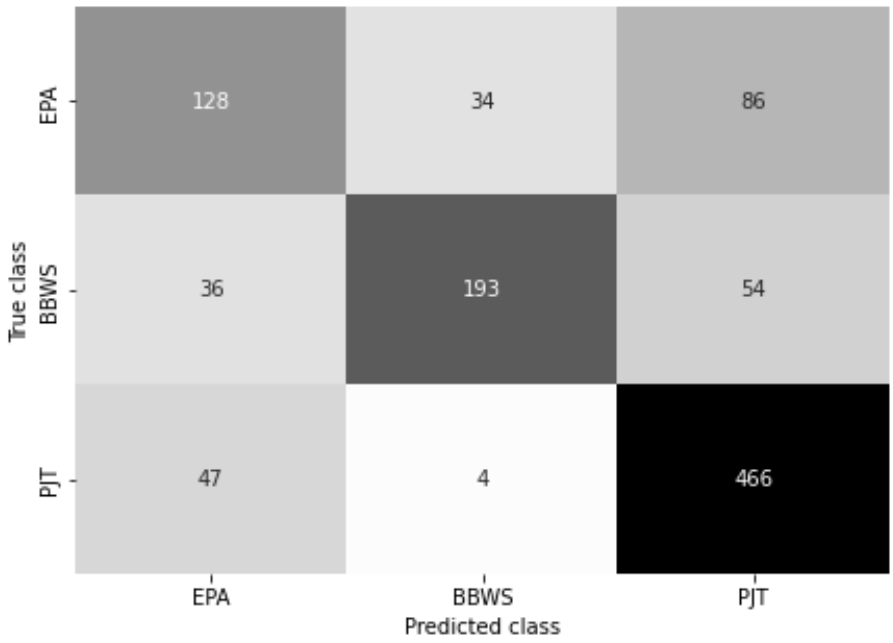

Figure 3.10: Confusion matrix of logistic regression.

To calculate the actual success rate of the logistic regression, we need to again account for the datapoints that are accidentally classified correctly. Instead of 8%, this time there is 24% that is

misclassified. This means that for approximately 36%, the class was guessed and thus only 64% of datapoints were successfully recognized as belonging to a specific author. The neural network which was 88% successful is thus a significant improvement over the logistic regression.

	precision	recall	f1-score	support
EPA	0.63	0.54	0.58	248
BBWS	0.85	0.69	0.76	283
PJT	0.77	0.90	0.83	517
accuracy			0.76	1048
macro avg	0.75	0.71	0.73	1048
weighted avg	0.76	0.76	0.75	1048

Table 3.3: Classification report of the logistic regression model.

4 DISCUSSION

In this discussion chapter the results and their implications are further analysed. This is done in four main sections. First off, an analysis of the main physical processes is provided. This is subsequently divided in four main elements that were recognised as important. The second section dives into the agencies and what the data tells us about their specific view on these processes. Section 4.3 then reviews the results of the neural network modelling that was done and what can be learned from that in the context of this study. The final section reviews the methodology that was used and the potential errors this could have caused.

4.1 PHYSICAL PROCESSES

4.1.1 Dissolved Oxygen

Dissolved oxygen is the most important water quality parameter in this system. This was first shown in the correlation matrix (figure 3.5), where oxygen showed high correlations with many of the other parameters. It also shows a clear pattern over the trajectory of the river (figure 3.1), i.e. high at the source, decent in the upper and middle sections and lower in the delta. Therefore it was not surprising to see oxygen show up as the dominant axis for the first principal component. That this holds true, not just for the overall PCA, but for the individual PCA's of the separate authorities, further stresses its importance.

According to the USGS, Water Science School (2018), oxygen enters stream mainly from the atmosphere, through groundwater discharge and through photosynthesis. It is then mainly influenced by organic matter and by temperature. Temperature in a tropic stream like the Brantas is fairly constant throughout the year and the main fluctuations occur near the source of the Brantas. The PCA does show an inverse relation between DO and temperature, but it is not particularly strong. Oxygen absorption from the atmosphere is driven by exposed surface area and by cascades or rapids. The exposed surface area is relatively high compared to the water volume in small streams. This is also where you tend to find natural cascades and rapids which thus explains the very high oxygen concentrations in the upper reaches. Oxygen from groundwater is likely not that significant for the Brantas, as according to the USGS, Water Science School (2018), this is mostly significant in streams with a large groundwater component.

Photosynthesis is a more likely influencer of oxygen concentrations in tropic streams, but this should be closely related to TSS concentrations as suspended solids make the water more turbid, reducing the light available for photosynthesis. TSS is however not relevant in PC1 and therefore we can assume that fluctuations in photosynthesis are not relevant to fluctuations in oxygen

concentrations, or at least at the reported scale as local and day-night fluctuations can not be accurately differentiated from this dataset. On the regional and monthly scales though, there is no significant influence on fluctuations by photosynthesis.

Organic matter, however, is a much more relevant component to the Brantas. This is seen in the loadings of PC1 for BOD and COD which are the main drivers in the oxygen process. Micro-organisms can lower dissolved oxygen levels through consumption of organic matter which is done in a BOD test (Boyles, 1997). COD is a more general measurement of the total oxidizable organic matter in a water sample. That COD correlates more with dissolved oxygen than BOD may look surprising at first, but could be explained by the fact that all BOD can naturally be consumed by dissolved oxygen whereas COD contains more biochemically inert components of organic matter. Especially in water that has not seen a recent influx of BOD, it is more likely that BOD is removed from the system through the consumption of oxygen. Hence, BOD concentrations go down as DO concentration goes down. The COD concentration could then tell you more about the original BOD content and pollution. Usually though, there is more BOD potential in the river than is actually being dissolved. A highly oxygenated river usually contains less BOD and vice versa.

4.1.2 Rainfall

Figure 3.3 showed that high concentrations in suspended solids mainly occurred during high flow periods, which we could also classify as the monsoon/rainy season. This relationship is confirmed by the USGS, Water Science School (2018) as a causality. Large amounts of rainfall typically bring with them large amounts of sediment, eroded soil and other debris. Increased flow is additionally less likely to allow for settlement of suspended solids.

Besides TSS, pH is also a major indicator of rainfall as was demonstrated by figure B.6 where pH too showed a seasonal component though perhaps not as clearly as TSS. This is likely due to the nature of rain being typically acidic (Liljestrand, 1985), especially compared to river water. This is caused by carbon dioxide concentrations in the air and can be further enhanced due to sulfuric and nitrogen oxides. That the lower pH of rainfall can then influence the pH of a river as well, was demonstrated by Lkr et al. (2022), Ling et al. (2017) and Ching et al. (2015).

That the second principal component resembles a rain related process can thus become clear from the loadings for pH and TSS. From section 3.1 and annex B it would seem that TSS is more strongly related to rainfall than pH, which would not become clear from PC2. This can be explained as a mathematical quirk, as pH is distributed more like a normal distribution and TSS is distributed more like a log-normal distribution. This causes pH to have relatively larger loadings in the PCA.

Other large loadings in PC2 are those for EC, BOD and oxygen. As rainwater is typically rather clean, it has a diluting effect on a water body and this is represented in the loadings for EC and BOD, but this reasoning does not make sense for oxygen, as rainfall usually contains a lot of oxygen. There exists a real correlation between dissolved oxygen and pH as was shown in figure 3.5. According to Makkaveev (2009) and Simonsen and Harremoës (1978), this relation is caused by the degradation of BOD, which consumes oxygen and releases carbondioxide, an acidic substance. Thus the degradation of BOD causes a drop in both DO and pH. As PC2 is focused more on pH than on TSS, it thus shows that it resembles not just rainfall, but this aspect

of pH as well. A margin of error has to be considered, however, as DO has a more dominant role in PC1, which is also a more important principal component, the loadings for dissolved oxygen in PC2 and subsequent PC's may also just be a mathematical fluke that is more representative of noise in the data than of actual physical processes (Björklund, 2019). Dissolved oxygen and TSS, for example, have pretty much no correlation (figure 3.5), but according to PC2 would be negatively correlated.

4.1.3 Surface waste runoff

In PC3, the main role is played by TSS. As concluded previously, suspended solids play a large role during high water levels and this thus hints at the fact that this too is a rainfall related process. Other loadings of significant magnitude here are found to be mostly BOD and COD. They are positively correlated with TSS whereas the relation in PC2, though weaker, was inverse. This all suggests that high BOD and COD loadings are not typically found during rainfall, but do occur under certain circumstances. Ching et al. (2015) found that in the Muar river in Malaysia:

> *Urban runoff during the flood events had increased loads of domestic wastes from the streets and sidewalks; nutrients from untreated sewage, discharges from agro-based industries, and manufacturing industries; leaves, grass clippings, and stormwater inputs from residential areas, all of which had increased the oxygen demand in the river.*

In short, all sorts of oxidizable pollutants can pile up on the surface and these are subsequently washed off by rainfall. This in turn concentrates the pollution load during rainfall periods. This relation is marked by the high loadings for BOD and COD in PC3. PC3 marks that with suspended solids, sometimes a lot of organic material is introduced as well.

4.1.4 The role of Electrical Conductivity

Electrical conductivity plays an interesting role in these principal components. It interacts with both PC1 and PC2, but it is more of a symptom of the processes than a driver. EC is mostly an indicator of the total amount of dissolved solids and therefore it can say something about the amount of pollution. We see that in PC1 it correlates strongly with BOD and COD, which is reinforced by the correlation coefficient in the correlation matrix (figure 3.5. In PC2, the rainfall process, it correlates negatively with total suspended solids. This indicates that during rainfall EC values drop. This is expected as rainfall is very scarce on dissolved minerals or other soluble substances. In PC3, which is also related to rainfall, it is absent, indicating that the increased BOD and COD loadings, which cause higher EC values, are offset by the increased volume of water.

In the cross-plot of EC versus COD (figure 3.6) an interesting relation between the two appeared. It seems that for average values of EC and COD, the two are mostly uncorrelated. However, for higher values of EC there is a large likelihood that COD values are also high. In reverse, large COD values tend to occur nearly exclusively when EC is also significantly above average. After investigating this a bit further, it was found that for an EC value of 1100 μS/cm, 29.0% of the COD measures were above 100 mg/L. From all measures off COD above 100 mg/L, 88.9%

reported an EC value above 1100 μS/cm. BOD shows a similar relation with EC, though it is less strong.

There is one major remark though about this relationship between EC and COD specifically. That has to do with the interference of chloride, as chloride increases EC and can cause higher readings for COD as well. For EC, this is not that important, as it is only used to determine the amount of Total Dissolved Solids and thus does not give a false reading, because chloride is just one of these ions comprising dissolved solids. Especially in brackish or salty waters it will be the main component that drives EC values. However, chloride can also interfere in COD testing (Saral and Goncaloğlu, 2008). This is a problem as chloride is not a component that can actually chemically oxidize oxygen and can thus artificially increase readings.

We can make an educated guess on the magnitude of this interference by looking at some of the larger EC values. At one location, Tambangan Tlocor, EC values reach into the 10000-30000 $\mu S/cm$ range. Other locations do not see values over 3000 $\mu S/cm$. In seawater, chloride is the dominant driver in EC values. Around the Indonesian islands the typical chloride concentration is ~20000 mg/L (Cotruvo, 2005) and EC is typically 50000 $\mu S/cm$ (Tyler et al., 2017). If we assume a linear relationship, then at Tlocor this would correspond to chloride values around 4000-12000 mg/L and at other locations this would not exceed 1200 mg/L. According to Saral and Goncaloğlu (2008) this would result in an additional COD measurement off 80-200 mg/L at Tlocor and less than 25 mg/L everywhere else. From the geography it becomes clear why Tlocor has such high EC values: it is very close to the sea and a likely target for seawater intrusion. COD values from this location will thus have to be treated with some scepticism as it is not clear whether chloride or organic pollution is the cause for increased COD levels.

In contrast, for other locations of the Brantas it is safe to assume that the ratio of chloride in the total dissolved solids is much lower than in seawater. The maximum influence of 25 mg/L COD deviation is thus probably an overestimation and hence it is safe to assume there is little to no influence of chloride on COD measurements at any other location.

As the disturbance of chloride in COD measurements is thus limited to a single location and the patterns between COD and EC are very similar to BOD and EC, we can safely assume that both exist. The significance of this lies in its practical applications. Electrical Conductivity is a parameter that can be measured in place and continuous, whereas COD takes a few hours in a laboratory and BOD even five days. Additionally, good EC readings are relatively cheap. This relationship in the Brantas could thus potentially be used for warning systems or targeted monitoring as well as expanding the monitoring network in general.

4.2 AGENCY PERSPECTIVES

4.2.1 BBWS

From the explained variance distribution of BBWS (figure 3.8, top left) it can be seen that the first principal component is relatively important and the remainder is distributed relatively flat between the others indicating that few other principal components carry much relevance. The first two principal components of BBWS are largely similar to the overall picture. In the first, dissolved oxygen plays a dominant role with significant negative loadings for BOD, COD and

EC. The second PC has a very large loading towards pH and a secondary loading for TSS. This indicates that it is slightly more oriented towards pH than to the overall picture of rainfall as TSS has a significantly smaller loading.

Larger differences occur in principal component 3. As opposed to TSS, it is now temperature that drives this principal component. Other large loadings are for dissolved oxygen and phosphates. This is not the surface waste runoff as was defined in section 4.1.3. The inverse loadings for temperature and oxygen make sense. As water temperature increases, the capacity to hold oxygen decreases. With BBWS measuring some of the larger fluctuations in temperature and dissolved oxygen, mostly in the very upstream section of the Brantas, it is somewhat expected that this would play a larger role for them. What is not expected here is the large negative loading for phosphate concentration. No credible correlation between phosphates and either temperature or dissolved oxygen can be found in any of the other graphs. As such it is thus uncertain if a real relationship exists or that it is just a mathematical quirk of the data.

The surface waste runoff where large suspended solids concentrations coincide with heavy BOD and COD loading is not found in the BBWS data. PC4 and PC5 both have large loadings for TSS, but they don't seem to agree on whether nitrites and nitrates have anything to do with that. As such it can safely be said that BBWS does not recognize this pattern of heavy river pollution through surface runoff.

4.2.2 EPA

Out of the three agencies, the EPA has the most explained variance captured in the first three principal components. This also means that out off these agencies, the EPA has data that describes these important processes in a significant way.

Their first principal component is again mostly driven by the oxygen process. Other large loadings are for BOD and COD while EC receives a smaller loading in PC1 when compared to the other agencies. Principal component 2 is mainly driven by TSS instead of pH. As TSS is more indicative of rainy seasons than pH, this PC2 is thus also more indicative of rainy seasons. PC2 is also relatively 'clean', i.e. few medium sized loadings for other parameters. This is useful, because it means those parameters have more significant meaning in subsequent PC's.

PC3 then describes the surface waste runoff. The main parameter is TSS, but it is nearly equally divided with DO, BOD and COD. With a positive correlation between the three, we can thus see that with the entry of suspended solids into the stream also comes a lot of organic waste, though usually not, as this would have otherwise been represented partially in PC2. The EPA does thus indeed recognize these as two separate processes.

4.2.3 PJT

The captured explained variance by PJT is relatively poor for the first principal component, but better for the second and third. PC4 through PC6 then also captures a fairly significant sum of explained variance with all three having nearly equal weights. Obviously though, the first three are more important.

The first principal component then, like the others is mainly driven by oxygen. Notable here is the role for EC. In contrast to the EPA, in PJT's data EC plays a major role in this oxygen process in which COD and BOD are the other key players. This indicates that PJT could potentially benefit the most from EC measurement campaigns as they are likely more indicative of high BOD or COD values for them.

On the second principal component the main role is played by TSS, which indicates that it is a rainfall related process. Other major loadings are for BOD and COD which correlate positively with suspended solids and for pH and nitrate which correlate negatively with TSS. The positive correlation between TSS and BOD/COD is an indication of the surface runoff process whereas the inverse correlation between TSS and pH is an indication of the regular rainfall process. That these two are combined in the same principal component shows that PJT sees the surface waste runoff as inherent to rainfall. As their measurement sites are located in the lower Brantas and Surabaya river branches, this indicates that the surface waste runoff occurs especially here in these urbanized areas.

4.3 NEURAL NETWORKS

The data classification by agency with neural networks achieved considerable success. After training on part of the dataset and validating with a second part to select the best performing iteration, when shown a new part of the dataset, still 92% of datapoints was selected to the right agency. Assuming a one in three accidental hit ratio, there was still 88% of datapoints that had a distinct enough signature for the neural network to detect the author. Compared to the logistic regression methods, which had a accuracy of 76% corresponding to a 64% success rate, this was a significant improvement. The logistic regression is a heavily simplified neural network with only a single note rather than a hundred. With 36 parameters rather than the 3003 used in the ANN, it can be compared to more traditional statistical methods for classifying data.

What was thus found is that much of the agency data is dissimilar enough to somewhat easily be recognized as belonging to one agency or another. For many of these datapoints there is thus a clear agency bias. When using the more sophisticated neural network classification system, it was found that even more data contained an agency signature. As this 88% of data is such a large portion of the total, most agency data contained a detectable signal. This signal could cause problems when somebody wants to use exact data from these agencies together.

In addition, because there is thus a clear agency signature, it has to be assumed that these signatures are created by some form of bias. Causes could be the locations or times measured at, the way samples are taken, methods used or man-made mistakes. There are many different possible causes of which it is difficult to determine what they are, let alone the magnitude. Therefore, I will refrain from delving deeper into this. It is also not relevant as the aim of this research is to determine the agency perspectives, not whose measurements are best. The results of the artificial neural network classification shows that some agencies are easier to classify then others, but this does not say anything about the independence of their data. Rather it shows that their data is slightly easier to recognize in comparison.

Other scientific work in which neural networks or machine learning was involved to find the author of a dataset could not be found. In order to ensure the validity of this methodology for

this purpose of author identification and its implications, this methodology and purpose was discussed with Taormina (2022).

4.4 METHODS

4.4.1 Data Editing and Imports

In between the collection of the data and its use in this report, a few steps were necessary to prepare the datasets for analysis. The datasets were unfortunately not delivered in a shape where they were ready for imports into Python and thus manual adjustments had to be made. These have been explained in the Data Treatment section (section 2.1) of the methodology chapter. It occurred several times that data points were somewhat to very ambiguous in what they actually meant. For example, a dash could mean a value of zero, it could mean something was undetected or it could mean that it was not measured at all. At other times mistakes were made with assigning data to the right column or row. This also needed fixing up. Because of all these modifications in the datafiles, a lot of interpretations on my end could have entered the datasets that could possibly alter data into something that was not originally recorded. There is no reason to believe this would exceed very limited cases only.

Whether or not this matters is a different question. Generally though, I find it hard to believe that during pre-processing large parts of the data have been modified in a significant enough way to impact later analyses. In terms of bias, one can argue that all datasets have been equal to the same form of bias, namely mine. Hence, we can therefore also argue that the same bias applies to all the data. As values in the data are not used as 'objectively true' and the focus lies on differences between agencies, an equal impact on all agencies should not create new differences and is thus not relevant. The important sidenote here is that the agencies were sometimes more or less neat in the way they stored their data and thus interference on my end is linked to the cleanliness of the datafiles. This could still have caused some relevant bias, though it should be clear from the methodology that editing happened in limited cases only and thus the influence can be seen as rather negligible.

There are thus a few ways in which errors or bias may have entered the datasets during data processing. The eventual analyses that were performed however treated the data as fairly robust and focus more on comparative figures and approximate figures then precise numbers. In comparison to the relatively minor changes that may or may not have occurred in data processing, the effect of these changes on the eventual analyses is thus pretty much negligible.

4.4.2 Principal Component Analysis

Principal component analysis is a mathematical simplification of a dataset, but it does rely on a few assumptions. The relations are linear and orthogonal to each other. This is directly derived from the mathematics that the principal components of a dataset X are represented by the eigenvectors of the covariance matrix $C = \frac{1}{n}XX^T$. This is by definition a linear and

orthonormal system (Shlens, 2014). In order to then get clear principal components, it relies on the data to have linear and orthogonal correlations.

The principal components are in essence best fits of the data in a multi-dimensional field. If data parameters that are related do not exercise a linear relation, the corresponding linear fit will then likely be a fairly poor fit. The second issue relates to the orthogonality of PCA. If the main processes that drive fluctuations in parameters are not orthogonal to each other, than PCA will be unable to find them and instead prescribe a best fit that is not necessarily on par with the principal component. Both of these issues were desribed by Shlens (2014).

As it cannot be assumed that these datasets have linear nor orthogonal processes behind them further testing of principal components is necessary. Jackson (1993) suggests several such methods, but two of the most simplistic ones are the Kaiser-Guttman method and the Broken-stick Method. The first assumes that an uncorrelated dataset would return equally large eigenvalues and the capture of explained variance by all principal components to be equal. For a PCA with 11 parameters, this would mean that every parameter that captures more than 9.09% of explained variance has a significant role. For our case, this would imply the first four principal components are significant, with the fourth only barely passing the bar. This method as Jackson (1993) explains is somewhat flawed though as random variance would cause at least half the principal components to be above average.

The second is the broken stick method, which divides the total explained variance in 11 (for our case) random parts. The first principal component then has to be larger than the largest part, etc. This would result in a criterion of 27.5% for PC1, 18.4% for PC2, 13.8% for PC3 and 10.8% for PC4. A principal component in this method can only be seen as significant if all previous components also cleared this bar. A keen eye would notice that the PCA conducted here does not meet the requirements of the Broken-stick model. (Jackson, 1993) noted however that the broken-stick criterion does have a tendency to underestimate the number of principal components. Seeing as the first three are still generally close to the broken stick and clearly larger than the Kaiser-Guttman criterion, the focus on these in this study can thus be justified.

It is therefore also paramount to not focus too much on the high loadings on parameters that already had previous higher loadings. Due to the issues with orthogonality, these are more likely to fit for noise than to present real correlations.

Nevertheless, these tests are much better than to select principal components on a preferred total capture of principal components, for example 90%. In our case, this would result in the need to explain 7 principal components, but that would barely support the goal of dimensionality reduction we set out to achieve. With 7 significant principal components we might as well have sticked to our original 11 parameters. This was already a fairly limited number of parameters to begin with though. (Dupont et al., 2020) mentions a study with 6 significant principal components, none of which capture more explained variance than the ones in this study, but the big difference is that it used 27 total parameters for PCA. As a result of that, those six PC's meet the broken-stick criterion. As such, there is a possibility this data's PCA would have met these requirements easier had I included more parameters.

4.4.3 Scaling Algorithm

Scaling is absolutely necessary for principal component analysis, as is the removal of outliers. The first is important so that all parameters have equal influence on PC loadings. If not, then principal components will be heavily biased towards those parameters with the numerically largest standard deviations and will provide little further information. The outlier removal, or temperance, is important because PCA is trying to optimize the squared distance from the best fitting line. It will thus be heavily biased towards those few datapoints that present a significant outlier. As could be seen from the distribution plot (figure 3.4) this dataset contains some pretty intense outliers, even on a logarithmic scale. However, simply removing any points with outliers would also remove a lot of information from the datasets, so I chose against removal. Instead I 'tempered' the value and thus the influence such outliers would have on PC's by setting them equal to the 1st and 99th percentile value of that parameter. By doing so I retained the information that this concerned a large value for this parameter, but without caring too much about how large the outlier actually was to prevent it from influencing the PCA too much.

Inbetween the 1st and 99th percentile a linear scaling was chosen. This does have the disadvantage of concentrating certain parameters close to the lower end value. Considering many parameters in the distribution plot (figure 3.4) showed a more linear relationship on a logarithmic axis, it can be argued that these parameters should be scaled logarithmically instead. In order to verify that the linear scaling was indeed the better option, this had to be tried as well.

The logarithmic scaling has been applied to the dataset on all parameters except for pH and temperature similarly to the distribution plot. Values were still clamped between the 1st and 99th percentile as the logarithmic values still contained outliers. This gave one problem with NH_3 in particular as its 1st percentile sits at 0 and the logarithmic value of this would be $-\infty$. PCA does not react well to values of ∞ or $-\infty$ and therefore these needed to be eliminated. Instead of the 1st percentile value for NH_3 the smallest numeric logarithmic value of NH_3 was chosen instead, which lies somewhere between the 1st and 2nd percentile.

The results of this logarithmic scaling can be found in annex C. While some of the loadings may now become more pronounced, the concentration of capture of explained variance is actually worse for this PCA both for the overall dataset as well as the individual agencies. The loadings did not provide a significantly easier explanation of the principal components either, so I decided to stick with the linear scaling instead.

5 CONCLUSIONS AND RECOMMENDATIONS

In this research I have explored the water quality data of three different agencies operating in the Brantas river. I specifically wanted to know how the measurements each agency took shaped their perspective on the Brantas' water quality. In addition, by combining the data of the agencies a more complete view of the Brantas was presented which the individual agencies are compared against.

This did come with a fair share of challenges. First of all, the data itself was difficult to read and formatted differently, not just across agencies, but within the datasets of agencies themselves too. To get all data in a single form for comparison, editing and interpretation was done on my end. Besides problems with data integrity, the methods themselves had limitations too. Scaling posed one of the more intricate challenges, as the more standard techniques were not adequate and therefore more subjective methods had to be chosen. Finally, there are limitations to the interpretation of the results, but there are still some important conclusions that can be made.

First of all, within the water quality parameters that were looked at, two to three important driving mechanisms can be identified. The first of these is related to the oxygen content of the river. The amount of oxygen demanded and present in the river ties in to many other parameters including temperature, ammonia and phosphate. Moreover, there is a high correlation between chemical oxygen demand and electrical conductivity, especially for higher values of the two. The second mechanism is induced by rainfall, which is an important driver for the pH and suspended solids in the river. As was found in section 3.1, high concentrations of suspended solids occurred mostly during wet seasons. Dry seasons were a near guarantee for low concentrations. Rainfall is also the driver behind a third mechanism, which describes what the suspended solids in the runoff are made of. In more natural regions, this is mostly sands and dust, but in urbanized regions all sorts of waste and pollution is swept of the streets and flushed into the river. This will cause an increase in oxygen demand, both chemical and biochemical.

While all authorities will find these two processes of oxygen and rain, the emphasis differs greatly. Especially the focus on solid waste varies wildly. The EPA distinguishes this rather well in PC3, apart from PC2. PJT, however, will find a single principal component that incorporates both aspects of rainfall runoff and BBWS does not identify it in any significant principal component. It was understood that these agencies take slightly different measurements, but the big differences in principal components was less expected. From the principal components, it becomes clear that with their measurements the agencies tell different stories of the river and its processes. As a consequence, authorities will develop different perspectives of the river water quality. All will notice the importance of oxygen, but only PJT and the EPA will notice the importance of surface waste runoff, whereas BBWS does not. The EPA recognizes the different types of runoff, but will not think of EC measurements as a powerful tool. PJT sees the negative

impact of rainfall runoff on the river water quality, but also sees it as inherent to rainfall rather than as a separate process.

As an indicator of how different the datasets are, neural networks were used to link the specific datapoints to the original author. This specific neural network found that 88% of data has a clear author signature. This was a significant improvement over the logistic regression which could only detect this in 64% of datapoints. The logistic regression is understood to find the more obvious differences, e.g. what you could also detect from the distribution plot of figure 3.4. The neural network however detects not just the obvious differences, but the subtle differences as well. Effectively, it proved how dissimilar the datasets from the agencies were. This dissimilarity of 88% underlines the conclusion found from PCA that with their measurements, these agencies write different stories of what they believe the river water quality to be.

5.1 RECOMMENDATIONS

PJT, BBWS and the EPA all have their stakes in the water quality management of the Brantas and hence they will run into each other when making policies. If these three are not in agreement on the state of the river, it is hard to believe they will reach an agreement on policies concerning it. Therefore it is paramount to share data to create mutual understanding. During this research, data from all three was collected and compared in a single dataset. This was not a straightforward task which came with several hurdles and obstacles. Therefore I have drawn up several recommendations to allow for better data sharing in the future.

The first and most important task is to add informative metadata to datafiles. The three essential pieces of information are time, location and method. Date was only consistently reported by PJT, which also recorded measurement time, though this was sometimes inexplicably absent. The EPA did often include the date on their measurements, but this was absent from a number of sheets as well. BBWS was least clear, as measures were taken quarterly and usually only the quarter of the measurement was recorded, which is not a very accurate recording of measurement time, as it was shown that many water quality parameters can easily change in that window. In order to compare similar measurements across different agencies, the date should be a minimum requirement and preferably time of day is also provided.

Location is another important piece of metadata that needs to be documented well. While all agencies did record the locations of their samples, it was usually more of a broad location and not very specific. Usually, only a name was provided, but this name was not typically easy to look up in resources like Google Maps. Additionally, not all three agencies used the same name for the same place and sometimes this even varied from year to year within an agency. GPS coordinates could be looked up though, sometimes within the document itself, sometimes in a secondary document. The problem here was not availability, though sometimes challenging, but the accuracy. Even though a large number of decimals was usually provided, when entering the coordinates in the satellite view of Google Maps it was not uncommon to find the location on land, tens of meters from the river. None of the authorities did particularly well on this front and therefore the actual location was still often an educated guess based on additional information such as the name of the bridge or the stream that it was measured in, the latter being especially relevant for the outflows of tributaries. One of the agencies that performed a bit

better on the location front was BBWS, who provided a secondary document detailing the name of the locations with coordinates, the stream measured, a color code to the type of location, an overview map depicting the gross location of that point and a small picture of what that location looked like on the ground. That last thing is surprisingly useful, even for someone working from miles away as this allows you to verify locations easily, if not in real life, then with a tool like Street View. What this document lacked was readability for computer programs. Every part of information that was extracted from this document had to be done by hand. Because of all the reasons stated above, copying and verifying the locations of measurement points was a very time consuming activity and this still resulted in only an approximate location. For true comparison of measurements with the same location, it is crucial to know not just on what bridge you took a measurement, but also if it was in the middle or on the side and if you took a sample from the top of the water or from a meter of depth. Even these small differences can make big differences in recorded values and currently there is no way to know.

The third part of metadata that was difficult to obtain were accurate descriptions of the methodology used to obtain samples and parameter values. Even if the same standard is used for measuring a certain parameter, differences can be made by the type of equipment and the way the equipment is calibrated or maintained. Just like with time and location it can be an important reason for why your readings are not the same as another agency.

Another major obstacle during this research was the poor digitization of data. While all data was at least digitized, i.e. there was no need to type readings from paper into Excel myself, the way in which this was done still largely resembled the paper form. This was annoying at best and problematic at worst. Problems ranged from excessive blank cells to temperatures being recorded as dates and everything in between. Most time consuming though was the constantly shifting structure. Because of this it was rarely ever possible to fully automate the import of data from multiple files into a single file. Every file and sometimes even every single sheet needed separate lines of code to import. Significant improvements can already be made if it is only kept in mind that files need to be computer readable and that the Excel sheet is not the end product.

BIBLIOGRAPHY

APHA (1992). Standard methods for the examination of water and wastewater, method 4500-no2. Technical report, American Public Health Association, American Water Works Association, Water Environment Federation. 18th Edition.

Atekwana, E. A., Atekwana, E. A., Rowe, R. S., Werkema, D. D., and Legall, F. D. (2004). The relationship of total dissolved solids measurements to bulk electrical conductivity in an aquifer contaminated with hydrocarbon. *Journal of Applied Geophysics*, 56(4):281–294.

BBWS (2020). Profil organisasi. https://sda.pu.go.id/balai/bbwsbrantas/page/profil/profil-organisasi. Retrieved: 5 May 2022.

Björklund, M. (2019). Be careful with your principal components. *Evolution*, 73(10):2151–2158.

Boyles, W. (1997). The science of chemical oxygen demand. *Technical information series, Booklet,(9)*, 24.

California Waterboards (2021). Guidance compendium for watershed monitoring and assessment: section 3.3.1 ammonia. SWAMP - Clean Water Team Citizen Monitoring Program.

Ching, Y. C., Lee, Y. H., Toriman, M. E., Abdullah, M., and Yatim, B. B. (2015). Effect of the big flood events on the water quality of the muar river, malaysia. *Sustainable Water Resources Management*, 1(2):97–110.

Cotruvo, J. A. (2005). Water Desalination Processes and Associated Health and Environmental Issues. *WATER CONDITIONING & PURIFICATION INTERNATIONAL MAGAZINE*.

Daims, H., Nielsen, J. L., Nielsen, P. H., Schleifer, K.-H., and Wagner, M. (2001). In situ characterization of nitrospira -like nitrite-oxidizing bacteria active in wastewater treatment plants. *Applied and Environmental Microbiology*, 67(11):5273–5284.

Dueñas, C. (2008). Water champion: Tjoek walujo subijanto blending corporate spirit with public service. https://web.archive.org/web/20081201093538/http://www.adb.org/Water/champions/subijanto.asp. Interview with Tjoek Walujo Subijanto, President Director Of PJT1. Retrieved: 5 May 2022.

Dupont, M. F., Elbourne, A., Cozzolino, D., Chapman, J., Truong, V. K., Crawford, R. J., and Latham, K. (2020). Chemometrics for environmental monitoring: a review. *Analytical Methods*, 12(38):4597–4620.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Hatzenpichler, R. (2012). Diversity, physiology, and niche differentiation of ammonia-oxidizing archaea. *Applied and Environmental Microbiology*, 78(21):7501–7510.

Heyes, A. (2000). None. *Journal of Regulatory Economics*, 17(2):107–129.

Houser, R. S. (2020). Pola 2020 pengelolaan sumber daya air wilayah sungai brantas. Unpublished translation document, Personal Communication.

Houser, R. S. (2021a). Examining patterns of plurality in water quality management in the brantas river basin. Draft 8 June 2021, TU Delft project team – updated 11 Oct, 2021.

Houser, R. S. (2021b). Excerpts from dlh jatim (east java epa) medium-term strategic plan, 2019-2024 (translated). Unpublished document, Personal Communication.

Jackson, D. A. (1993). Stopping rules in principal components analysis: A comparison of heuristical and statistical approaches. *Ecology*, 74(8):2204–2214.

Junier, S. (2017). *Modelling expertise: Experts and expertise in the implementation of the Water Framework Directive in the Netherlands*. PhD thesis, Delft University of Technology. https://doi.org/10.4233/uuid:eea8a911-f786-4158-a67e-b99663275bf8.

Li, D., Xu, X., Li, Z., Wang, T., and Wang, C. (2020). Detection methods of ammonia nitrogen in water: A review. *TrAC Trends in Analytical Chemistry*, 127:115890.

Liljestrand, H. M. (1985). Average rainwater ph, concepts of atmospheric acidity, and buffering in open systems. *Atmospheric Environment (1967)*, 19(3):487–499.

Ling, T.-Y., Soo, C.-L., Liew, J.-J., Nyanti, L., Sim, S.-F., and Grinang, J. (2017). Influence of rainfall on the physicochemical characteristics of a tropical river in sarawak, malaysia. *Polish Journal of Environmental Studies*, 26(5):2053–2065.

Lkr, A., Singh, M. R., and Puro, N. (2022). Spatio-temporal influence on river water chemistry of doyang river, nagaland, india, using multivariate techniques. *International Journal of Environmental Science and Technology*.

Makarim, N., Ridho, R., Sarjanto, A., Salim, A., Setiawan, M. A., Ratunanda, D., Wawointana, F., Dahlan, R., Afsah, S., Laplante, B., and Wheeler, D. (1995). What is proper? reputational incentives for pollution control in indonesia. Technical report, World Bank.

Makkaveev, P. N. (2009). The features of the correlation between the ph values and the dissolved oxygen at the chistaya balka test area in the northern caspian sea. *Oceanology*, 49(4):466–472.

Ou, G. and Murphey, Y. L. (2007). Multi-class pattern classification using neural networks. *Pattern Recognition*, 40(1):4–18.

Partnership, M. W. W. (2022). ph and alkalinity. https://www.umass.edu/mwwp/protocols/rivers/ph_alkalinity_river.html. Retrieved on 22 April 2022.

PJT1 (2022). Perum jasa tirta i website. https://jasatirta1.co.id/, Retrieved: 9 May 2022.

Pérez-Bendito, D. and Rubio, S. (1999). Chapter 14 chemometrics in environmental analysis. In *Environmental Analytical Chemistry*, volume 32 of *Comprehensive Analytical Chemistry*, pages 747–833. Elsevier.

Saral, A. and Goncaloğlu, B. I. (2008). Determination of real cod in highly chlorinated wastewaters. *CLEAN – Soil, Air, Water*, 36(12):996–1000.

Shlens, J. (2014). A tutorial on principal component analysis.

Simonsen, J. and Harremoës, P. (1978). Oxygen and ph fluctuations in rivers. *Water Research*, 12(7):477–489.

Taormina, R. (2022). private communication.

Tyler, R. H., Boyer, T. P., Minami, T., Zweng, M. M., and Reagan, J. R. (2017). Electrical conductivity of the global ocean. *Earth, Planets and Space*, 69(1).

United States Environmental Protection Agency (2012a). Volunteer Stream Monitoring: A Methods Manual, Chapter 5.6 Phoshorus. Retrieved From: https://archive.epa.gov/water/archive/web/html/vms56.html.

United States Environmental Protection Agency (2012b). Volunteer Stream Monitoring: A Methods Manual, Chapter 5.7 Nitrates. Retrieved From: https://archive.epa.gov/water/archive/web/html/vms57.html.

United States Environmental Protection Agency (2012c). Volunteer Stream Monitoring: A Methods Manual, Chapter 5.8 Total Solids. Retrieved From: https://archive.epa.gov/water/archive/web/html/vms58.html.

USGS, Water Science School (2018). Water Quality Information by Topic. https://www.usgs.gov/special-topics/water-science-school/science/water-quality-information-topic#overview. Last accessed on 2022-02-08.

Valiant, R. (2013). Integrated water resources management: Water for multipurpose, experience in the brantas river basin, east java. Retrieved from: https://www.academia.edu/7423617/Integrated_Water_Resources_Management_in_the_Brantas_River_Basin_East_Java_Indonesia on 9 May 2022.

Van Breukelen, B. M. (2022). private communication.

Visser, S. A. (2019). Study into the short and long term (re)production of relations between communities, inorganic solid waste and the surabaya river, indonesia. Master's thesis, Delft University of Technology.

Wood, M. S. (2014). Estimating suspended sediment in rivers using acoustic doppler meters. Technical report, USGS.

A MEASUREMENT LOCATIONS

ID	Location	Lat	Lon	EPA	BBWS	PJT
1	Arboretrum	-7.754144	112.526465	y	y	-
2	J. Brantas	-7.864163	112.526695	-	y	-
3	J. Pendem	-7.902844	112.574287	y	y	-
4	J. Dinoyo	-7.939556	112.611917	y	-	-
5	J. Gadang	-8.024444	112.632889	y	-	-
6	J. Sengguruh	-8.182012	112.546483	y	y	-
7	J. Karangkates	-8.156556	112.434333	y	-	-
8	J. Brawijaya	-8.171410	112.335754	-	y	-
9	J. Selopuro	-8.166111	112.300556	y	-	-
10	J. Glondong	-8.153333	112.217500	y	-	-
11	J. Trisula	-8.139931	112.146111	y	y	-
12	T. Ngunut II	-8.069417	111.983389	y	-	-
13	J. Ngujang	-8.016528	111.925361	y	-	-
14	T. Maesan	-7.942893	111.953021	-	y	-
15	J. Masjid Agung	-7.827500	112.009167	y	-	-
16	J. Jong Biru	-7.782288	112.008611	y	y	-
17	J. Papar	-7.694722	112.075278	y	-	-
18	J. Lama Kertosono	-7.600330	112.108677	y	y	-
19	T. Ngrombot	-7.549491	112.122120	-	y	y
20	J. Ploso	-7.460857	112.224536	y	y	y
21	T. Cheil Jedang	-7.446067	112.249833	-	y	-
22	T. Betro Kemlagi	-7.457067	112.338150	-	y	-
23	J. Les Padangan	-7.459846	112.432012	y	y	y
24	T. Canggu	-7.436230	112.459880	-	-	y
25	J. Jetis Permai	-7.427729	112.473867	y	y	-
26	J. Perning	-7.409549	112.492002	y	-	y
27	J. Legundi	-7.387316	112.577057	y	y	y
28	T. Cangkir	-7.363910	112.636077	y	-	y
29	T. Bambe	-7.351130	112.663979	y	-	y
30	J. Sepanjang	-7.344907	112.691618	-	y	y
31	J. Karang Pilang	-7.343333	112.695167	y	-	-
32	B. Gunungsari	-7.308404	112.718669	y	-	y

Table A.1: Measurement locations in the Brantas. Coordinates are the average of those provided by agencies. Columns 4-6 mark which agencies have measurements at that location with 'y'.

ID	Location	Lat	Lon	EPA	BBWS	PJT
33	J. Joyoboyo	-7.299715	112.736669	y	y	-
34	J. Bungkuk Ngagel	-7.296667	112.741667	-	y	-
35	J. Bung Tomo	-7.288667	112.744167	y	-	-
36	J. Sono Kembang	-7.272611	112.744750	y	-	-
37	J. Yos Sudarso	-7.262376	112.746109	-	y	-
38	J. Pasar Besar	-7.247806	112.742167	y	-	-
39	J. Petekan	-7.222408	112.738004	y	y	y
40	J. Karangrejo	-8.003669	111.911453	-	y	-
41	J. Begedheng	-7.506564	112.149615	-	y	-
42	J. Munung	-7.489369	112.168964	-	y	-
43	J. Pulorejo	-7.462186	112.428411	-	y	-
44	J. Jetis IV	-7.410991	112.473279	-	y	y
45	Muara K. Kwangen	-7.406350	112.486127	-	y	-
46	Hulu Kali Tengah	-7.365750	112.603472	y	-	-
47	WWG Kali Tengah	-7.359917	112.627861	y	-	-
48	J. Bambe	-7.351351	112.661987	y	y	y
49	Muara K. Kedurus	-7.307180	112.720310	-	-	y
50	J. Ciro	-7.426410	112.478400	-	-	y
51	J. Ngoro	-7.494470	112.563905	-	y	-
52	J. Carat	-7.562654	112.684461	-	y	y
53	J. Wonokromo	-7.300320	112.739970	-	y	-
54	J. Nginden Intan	-7.307754	112.768247	-	y	-
55	J. Merr II	-7.310556	112.780556	y	y	-
56	T. Wonorejo	-7.308003	112.798863	y	y	-
57	J. By Pass Mojokerto	-7.445140	112.459201	y	y	-
58	B. Lengkong Baru	-7.445160	112.466050	-	-	y
59	J. Ngrame II	-7.475615	112.560213	y	-	-
60	J. Porong	-7.545903	112.698162	y	y	y
61	T. Tlocor	-7.545848	112.819300	y	y	-
62	I. Karang Pilang	-7.348668	112.680082	y	-	y
63	I. Ngagel	-7.300350	112.741300	-	-	y
64	I. Pelayaran	-7.409350	112.529450	-	-	y
65	I. Tawangsari	-7.351940	112.676110	-	-	y

Table A.2: Measurement Locations in the Brantas. Coordinates are the average from those provided by agencies. Columns 4-6 mark which agencies have measurements at that location with 'y'.

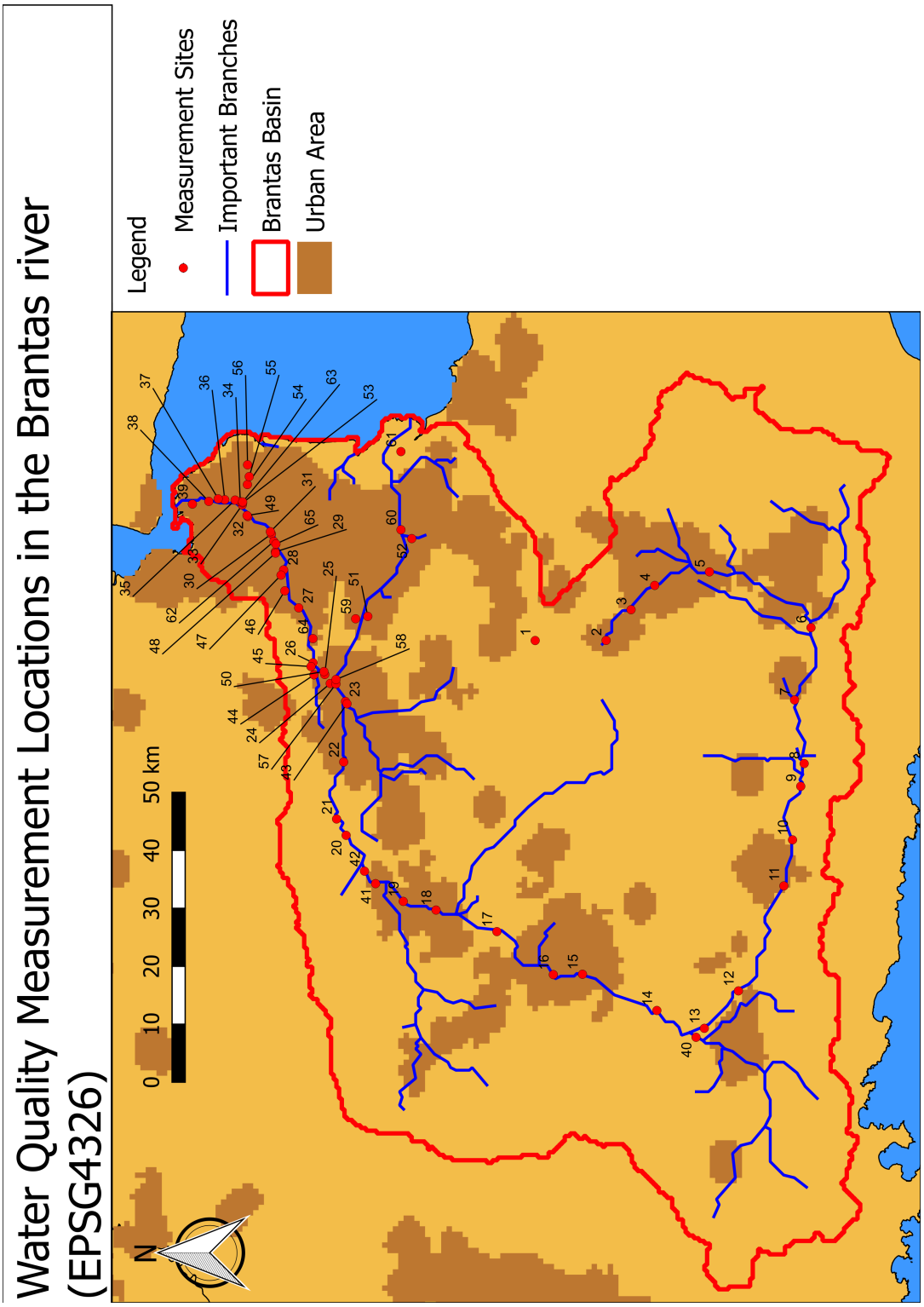

Figure A.1: Measurement Locations in the Brantas. Numbers correspond to locations as seen in tables of this annex and other figures such as in the diagrams of annex B.

B PARAMETER BOXPLOTS

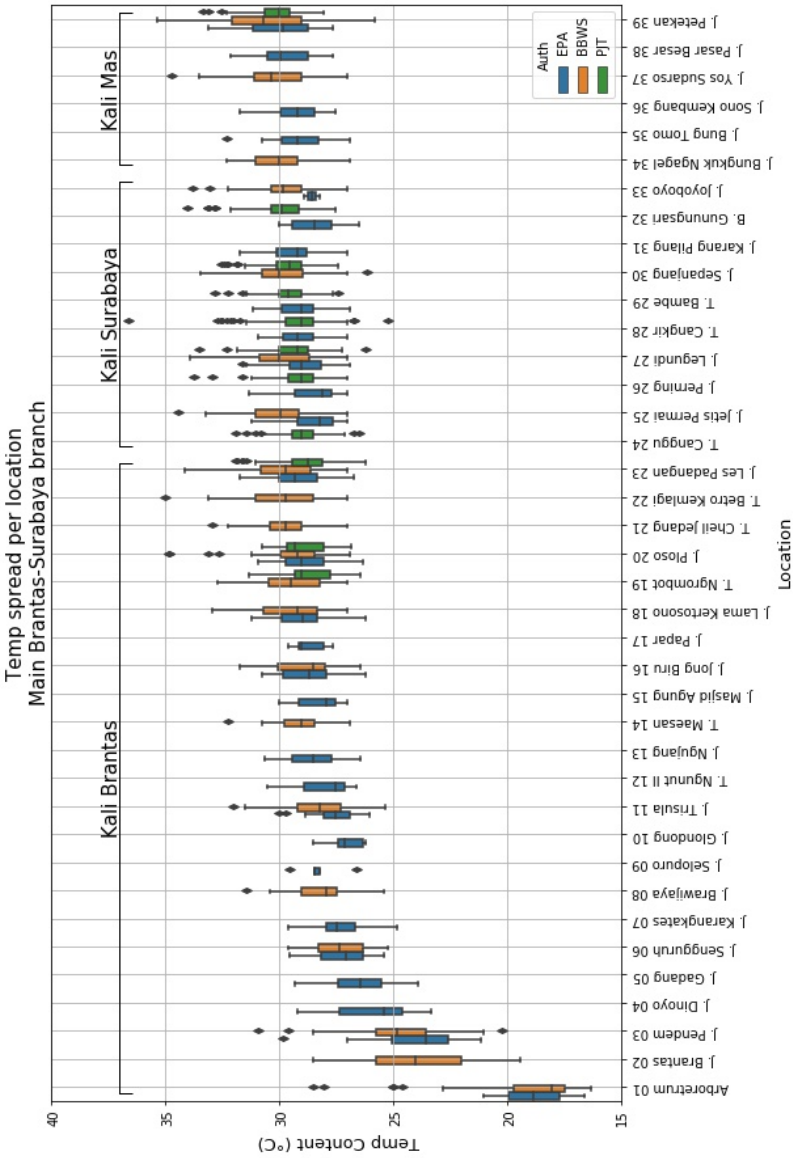

Figure B.1: Water temperature along the main Brantas stretch.

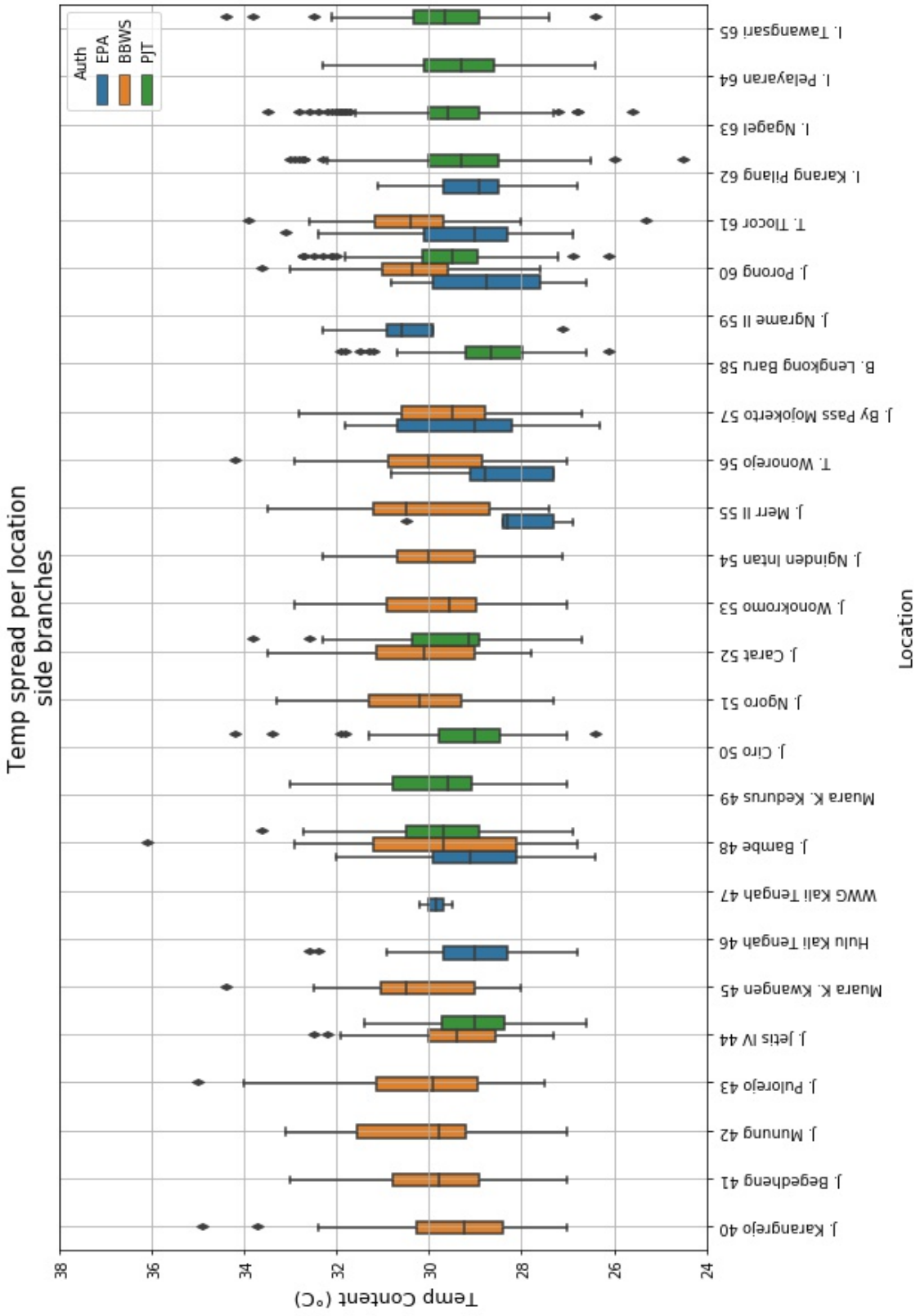

Figure B.2: Water temperature in the branches of the Brantas.

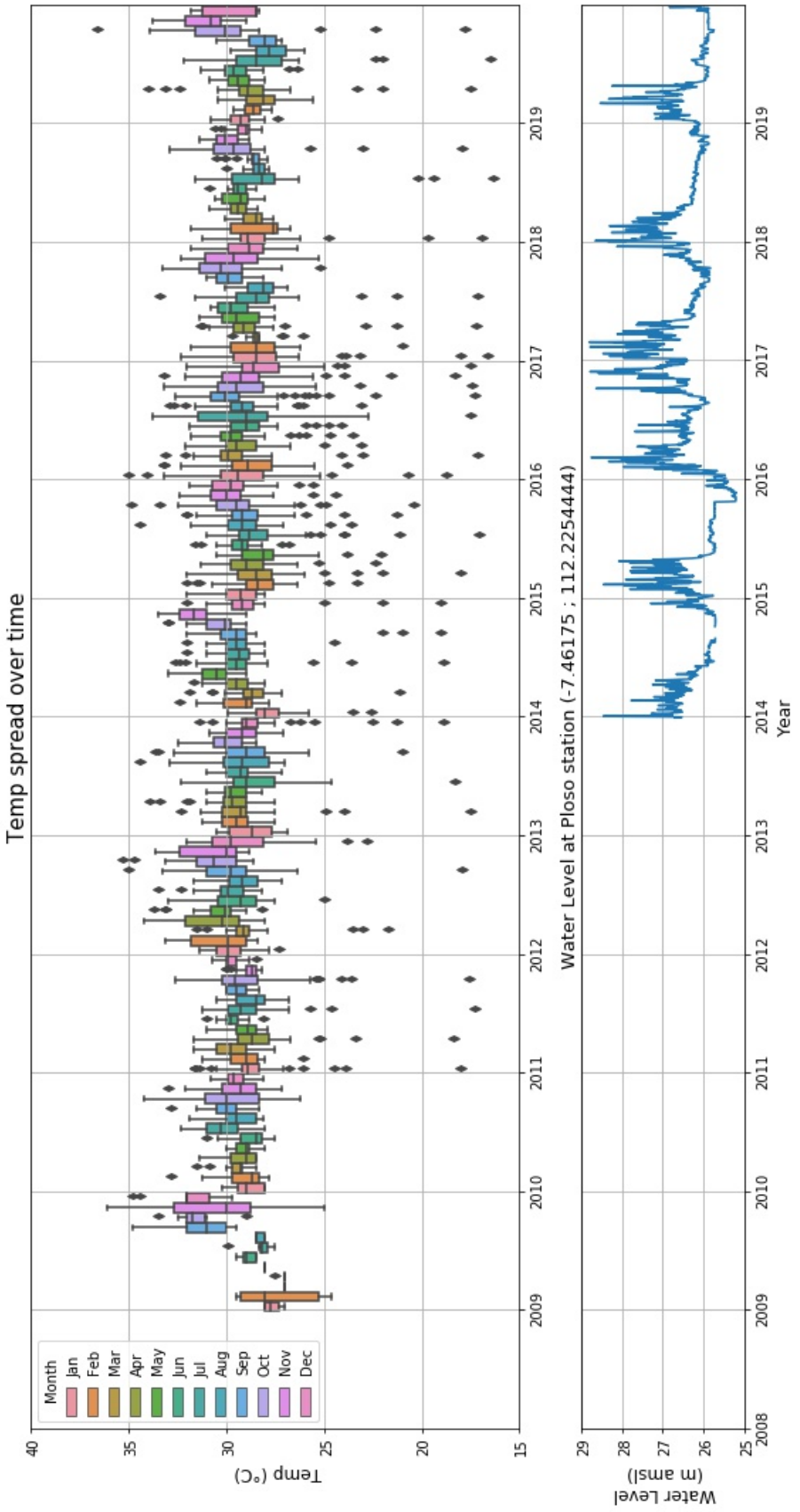

Figure B.3: Water temperature in the Brantas over time.

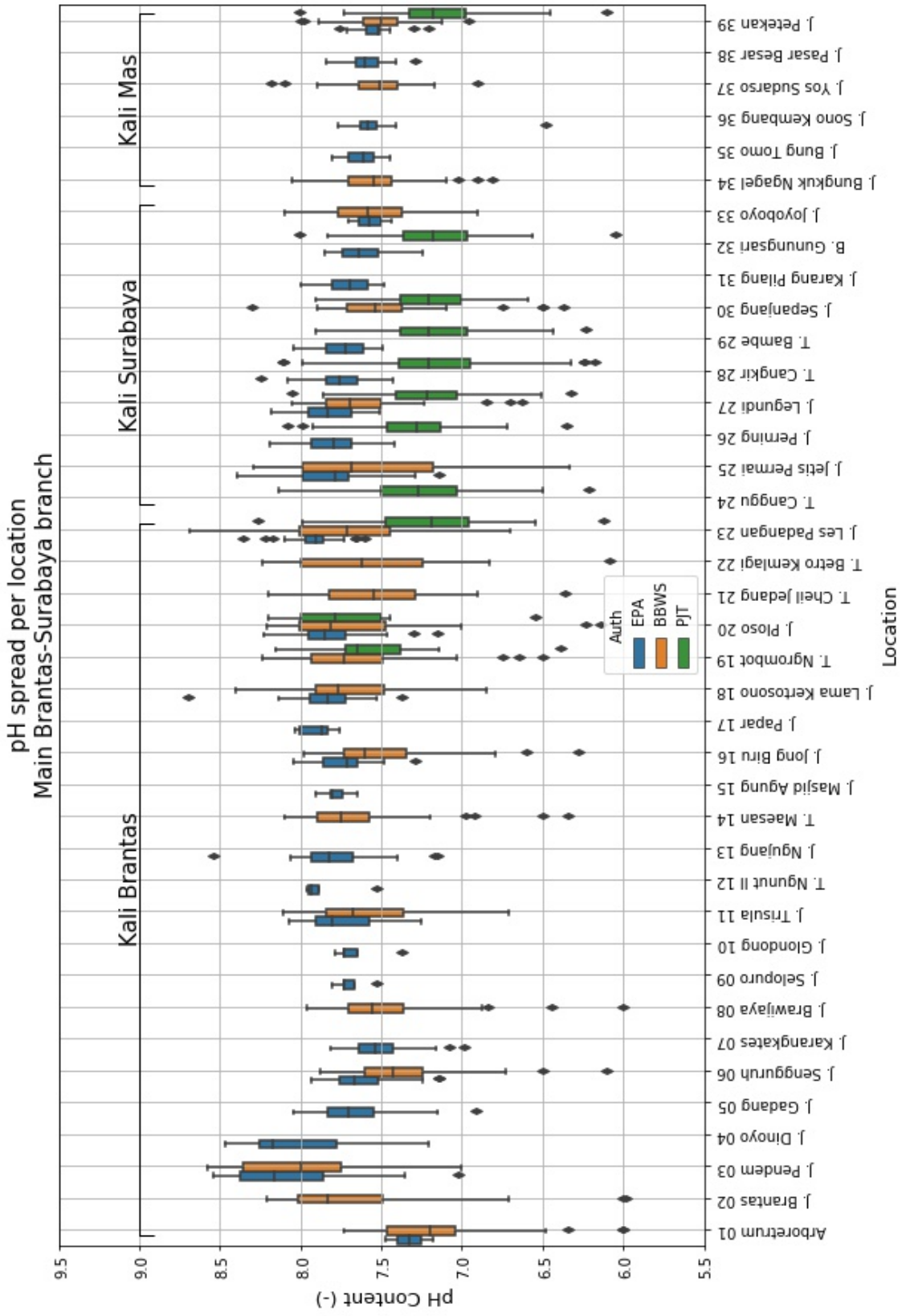

Figure B.4: pH values along the main Brantas stretch.

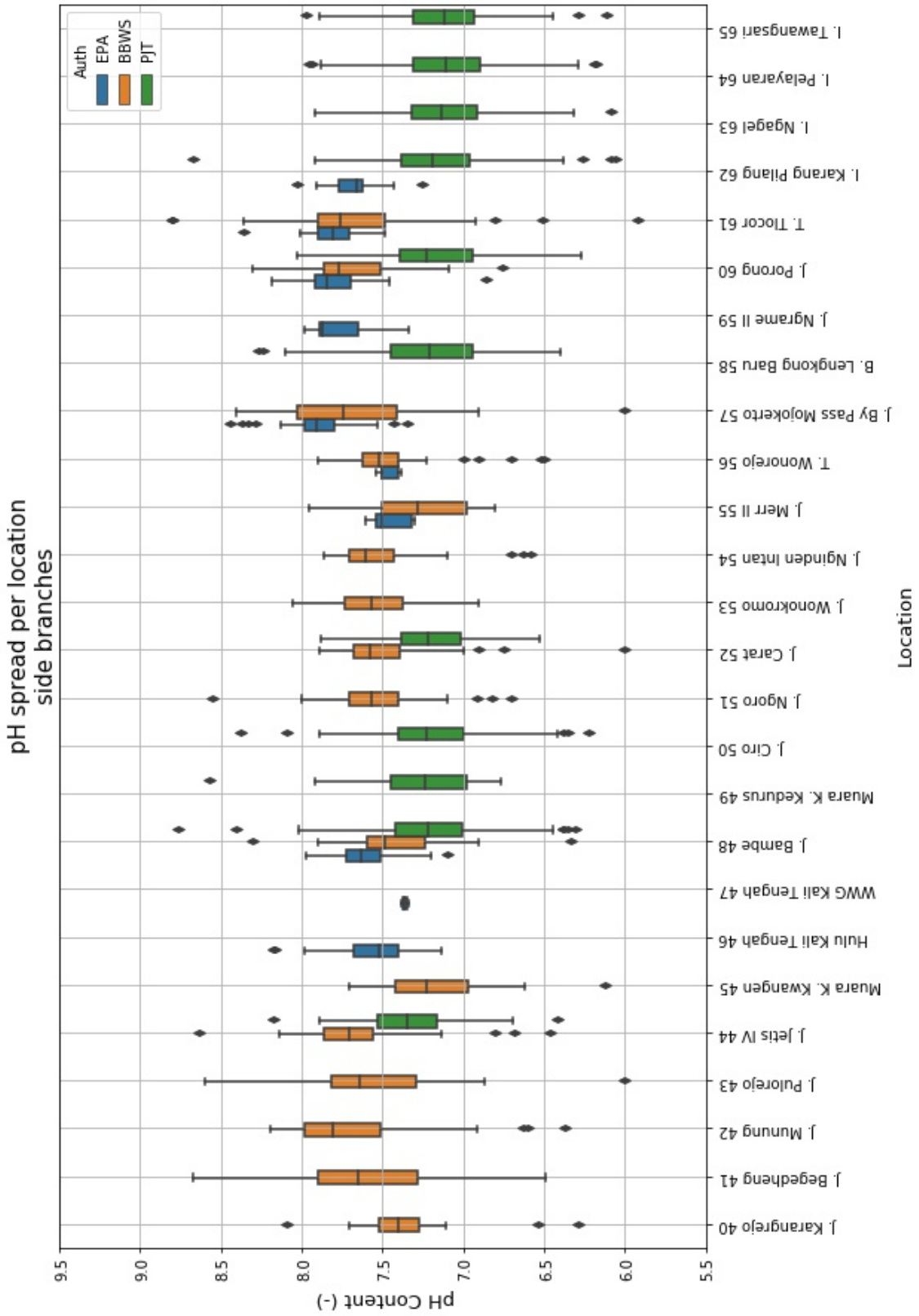

Figure B.5: pH values in the branches of the Brantas.

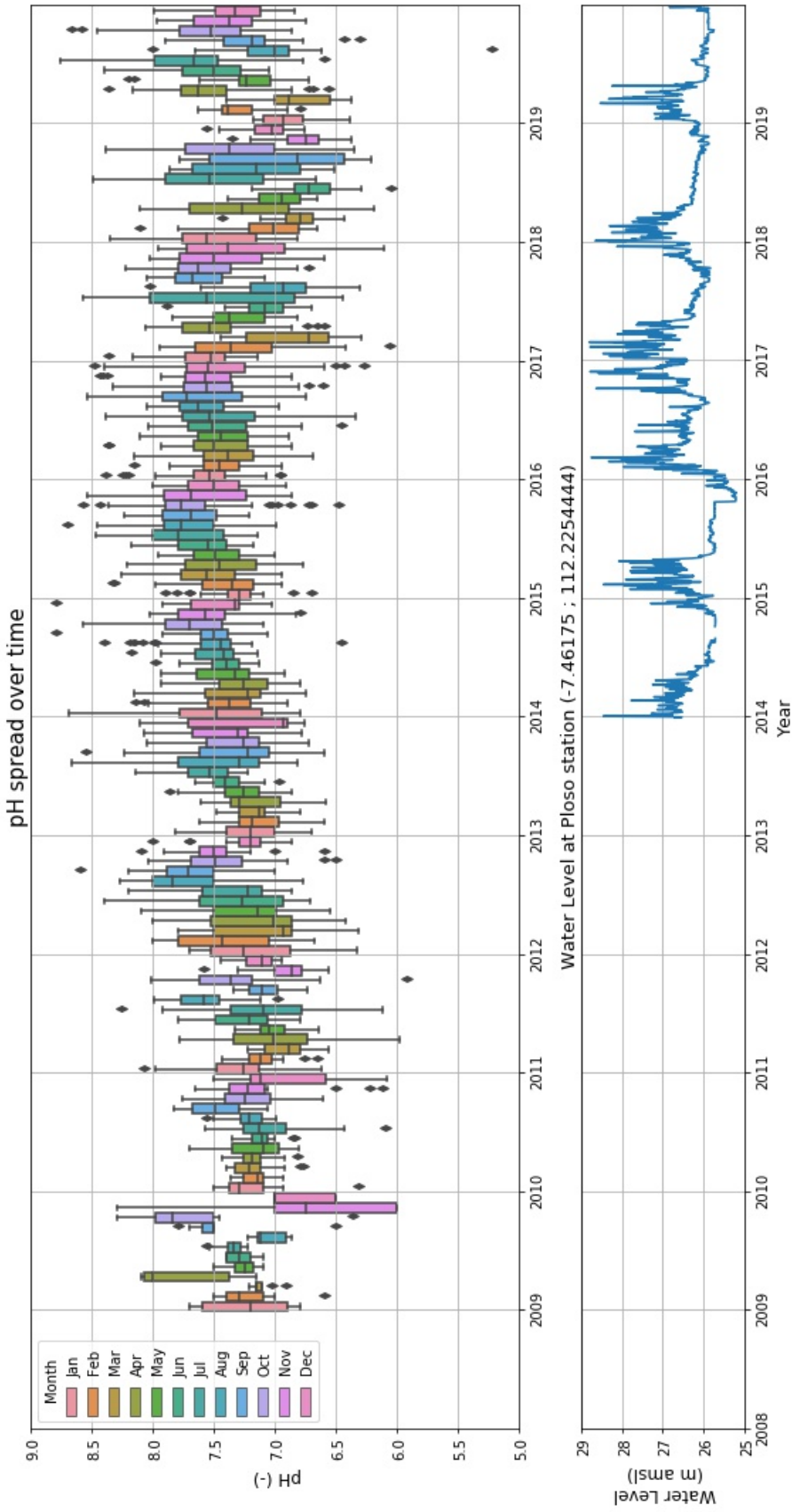

Figure B.6: pH values in the Brantas over time.

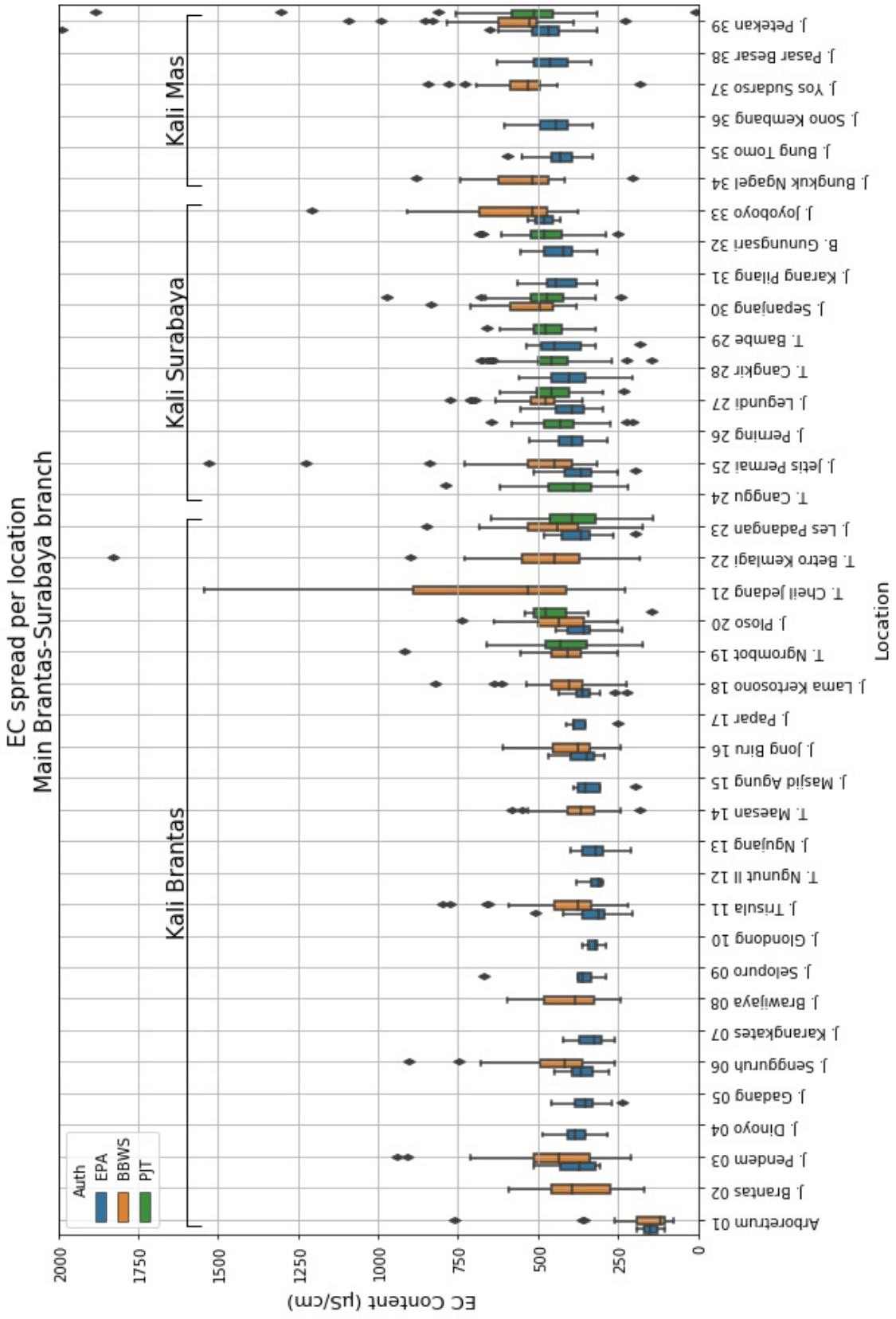

Figure B.7: EC values along the main Brantas stretch.

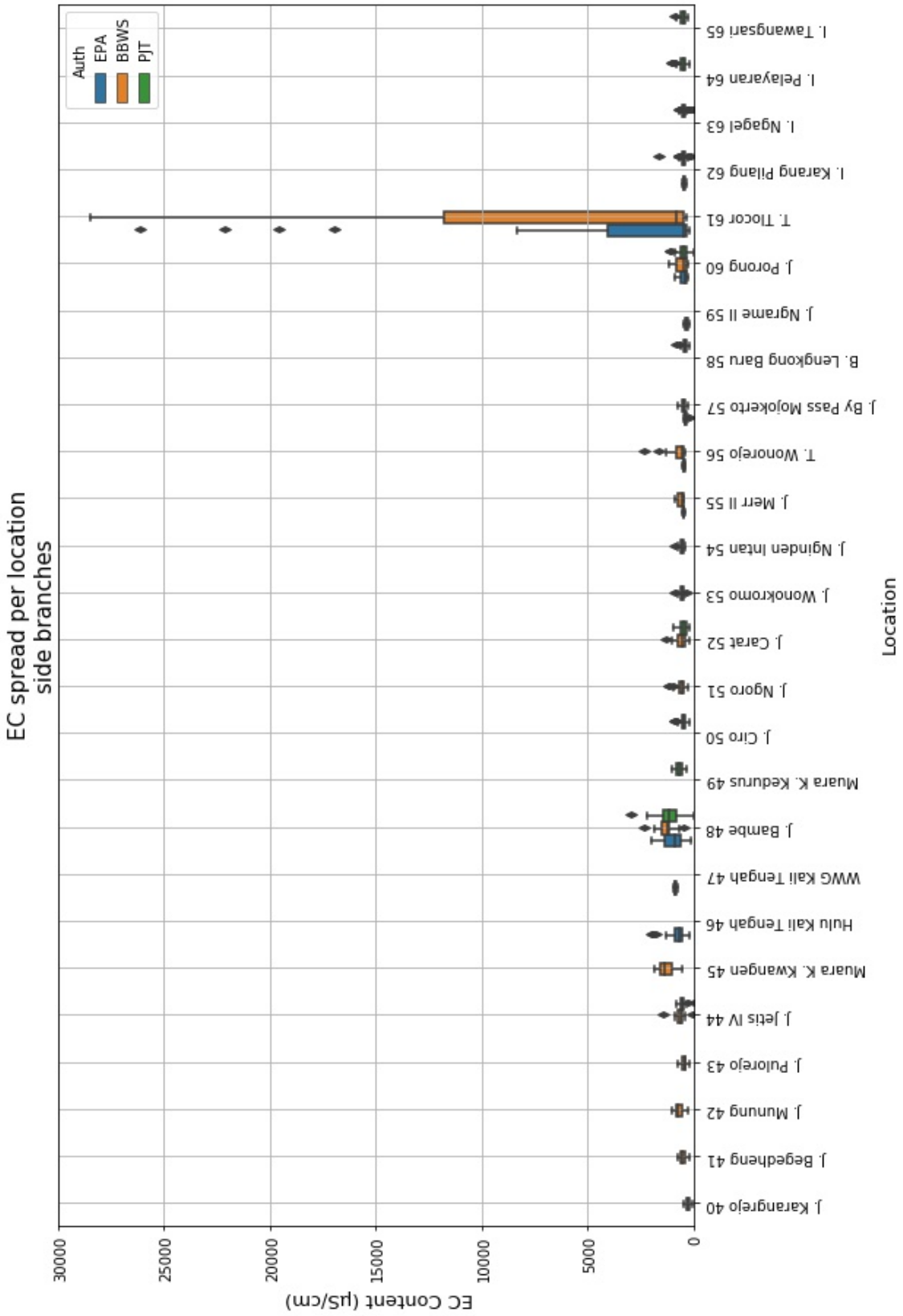

Figure B.8: EC values in the branches of the Brantas.

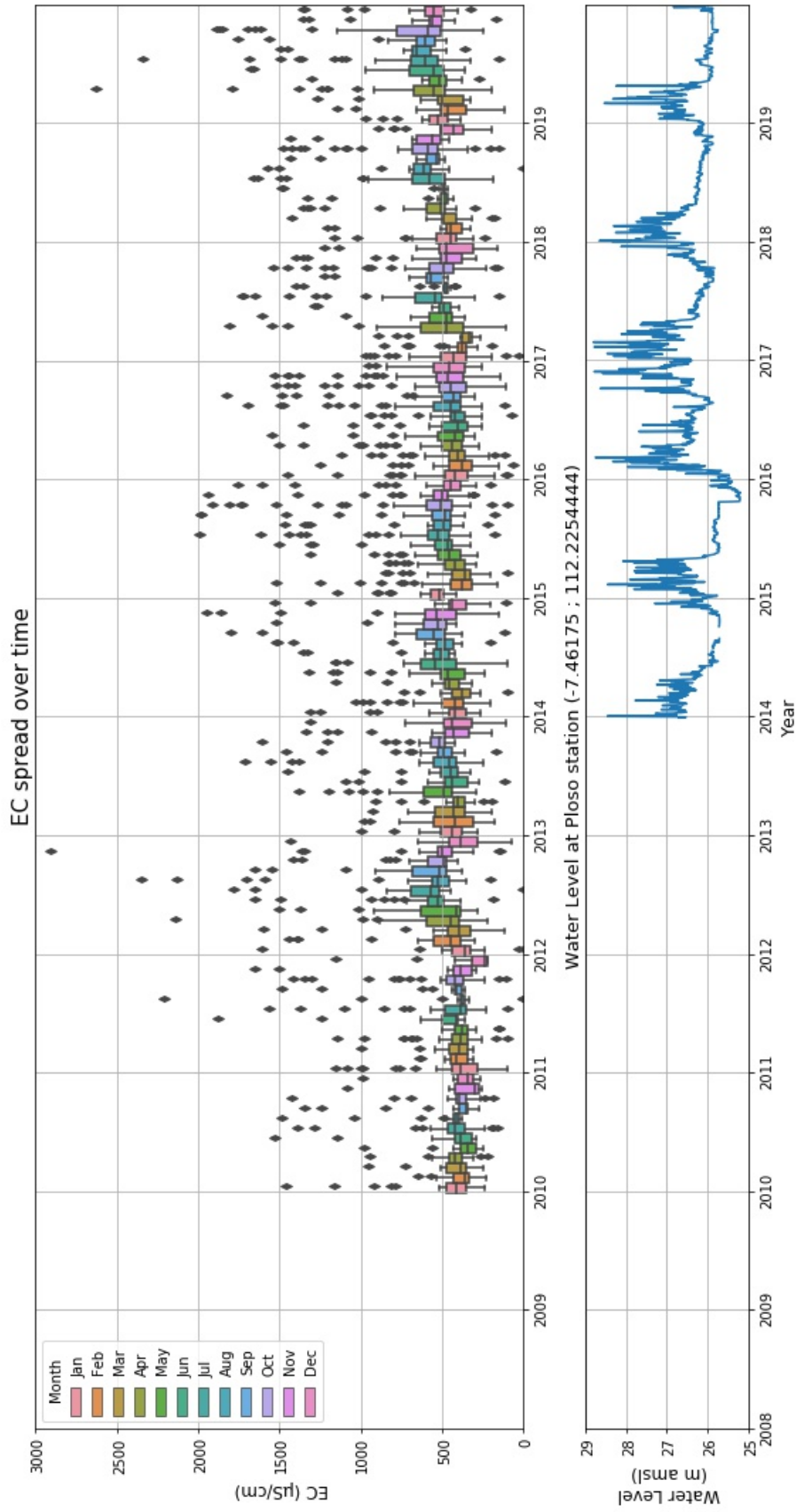

Figure B.9: EC values in the Brantas over time.

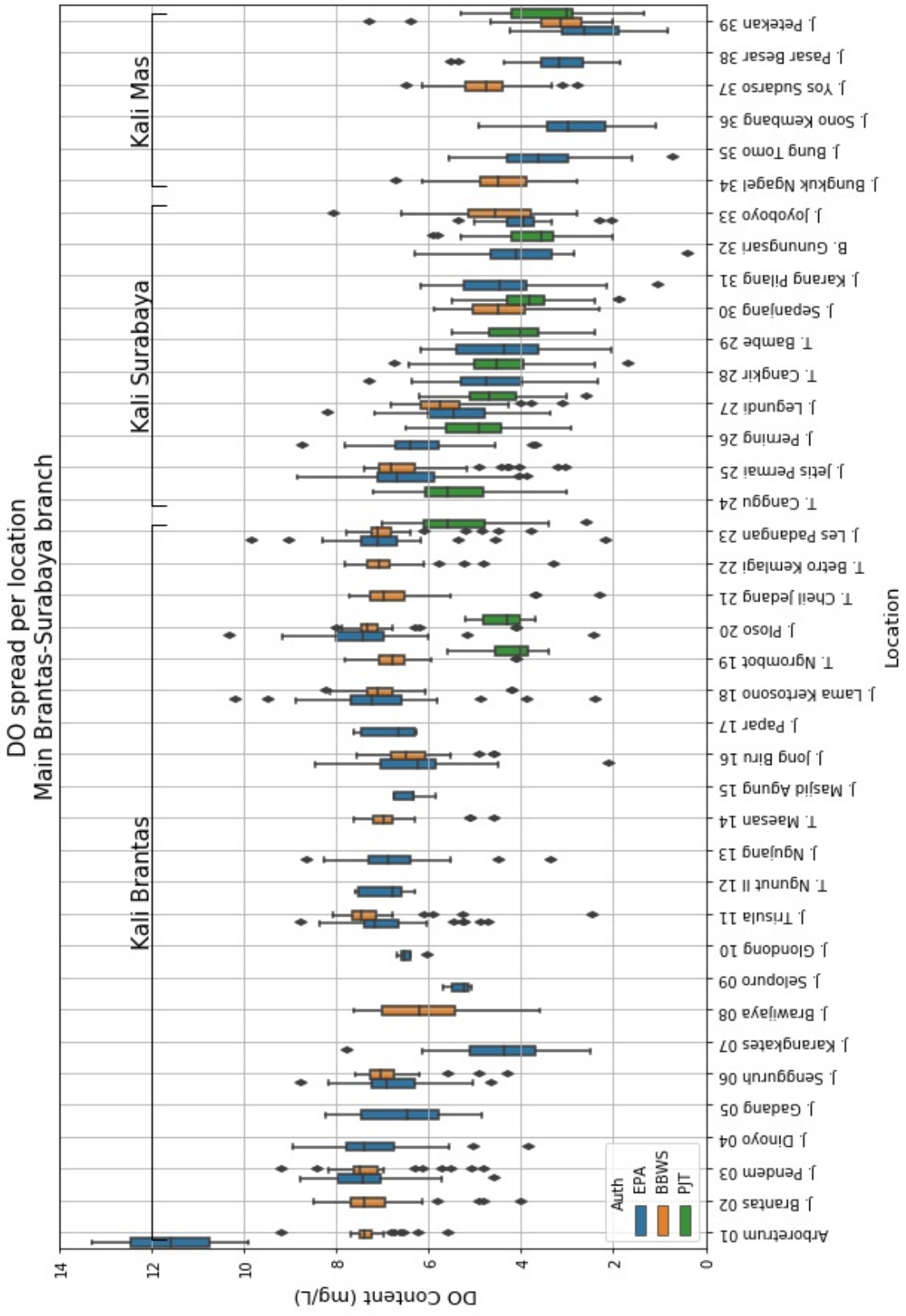

Figure B.10: Dissolved oxygen concentrations along the main Brantas stretch.

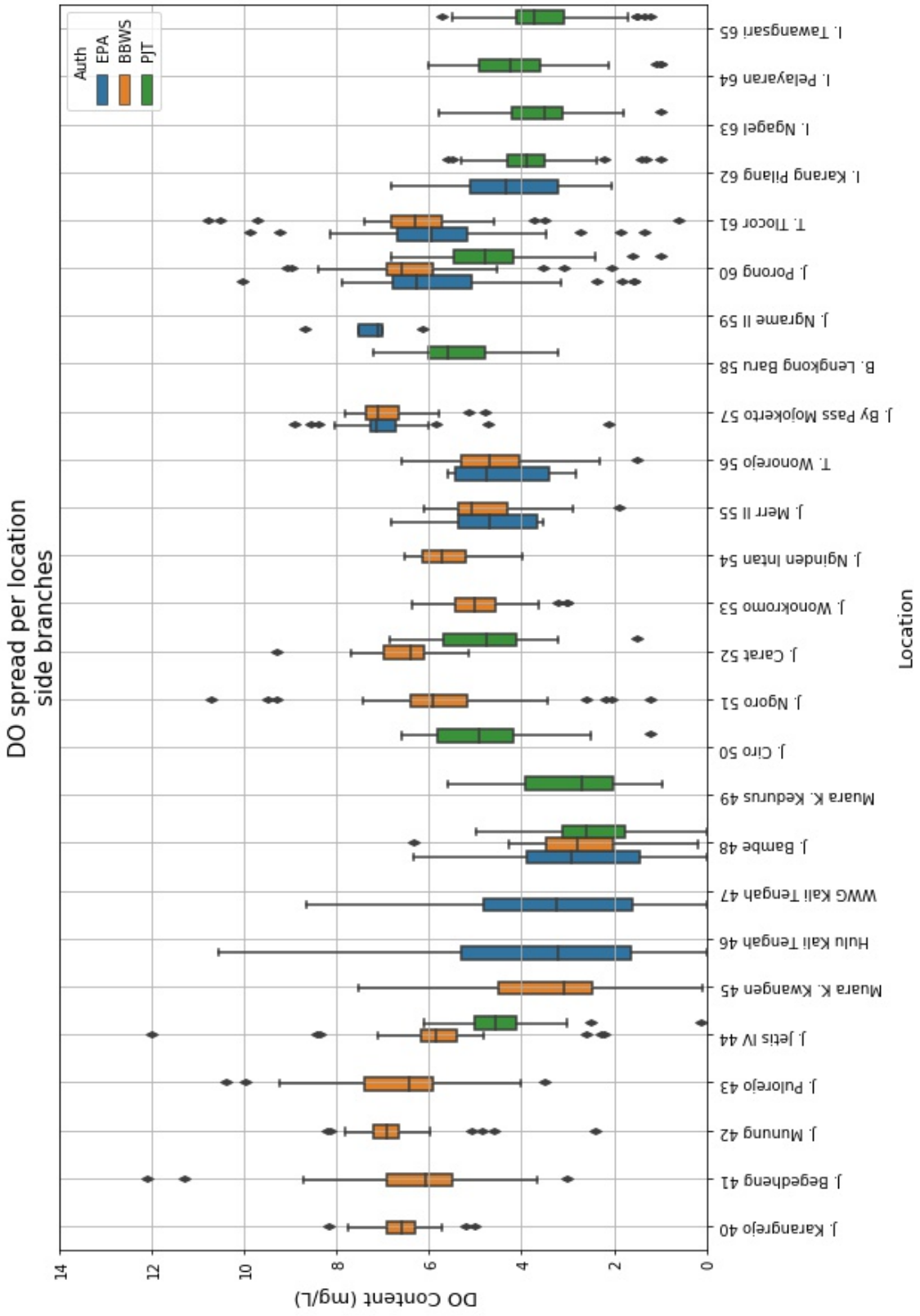

Figure B.11: Dissolved oxygen concentrations in the branches of the Brantas.

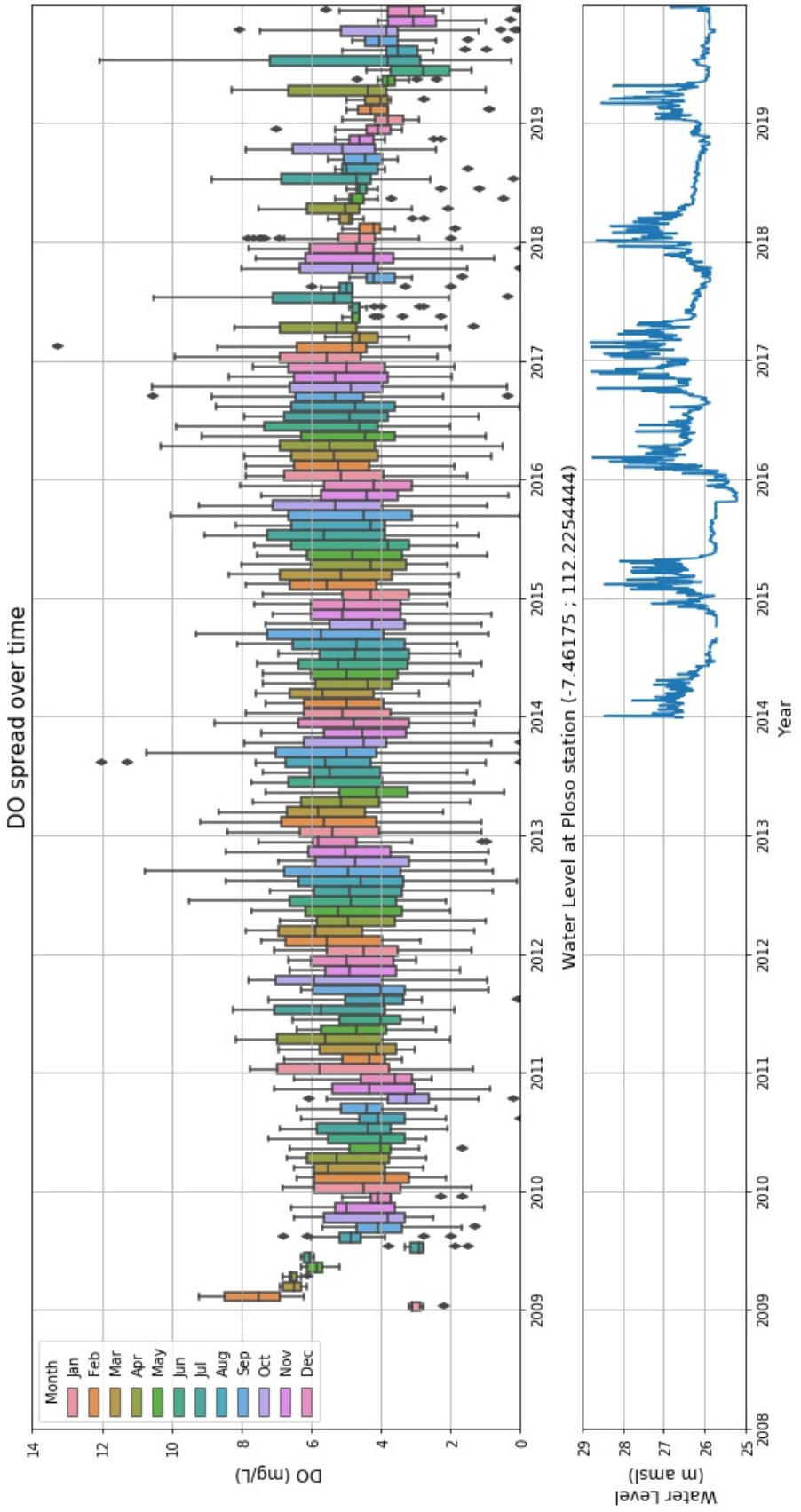

Figure B.12: Dissolved oxygen concentrations in the Brantas over time.

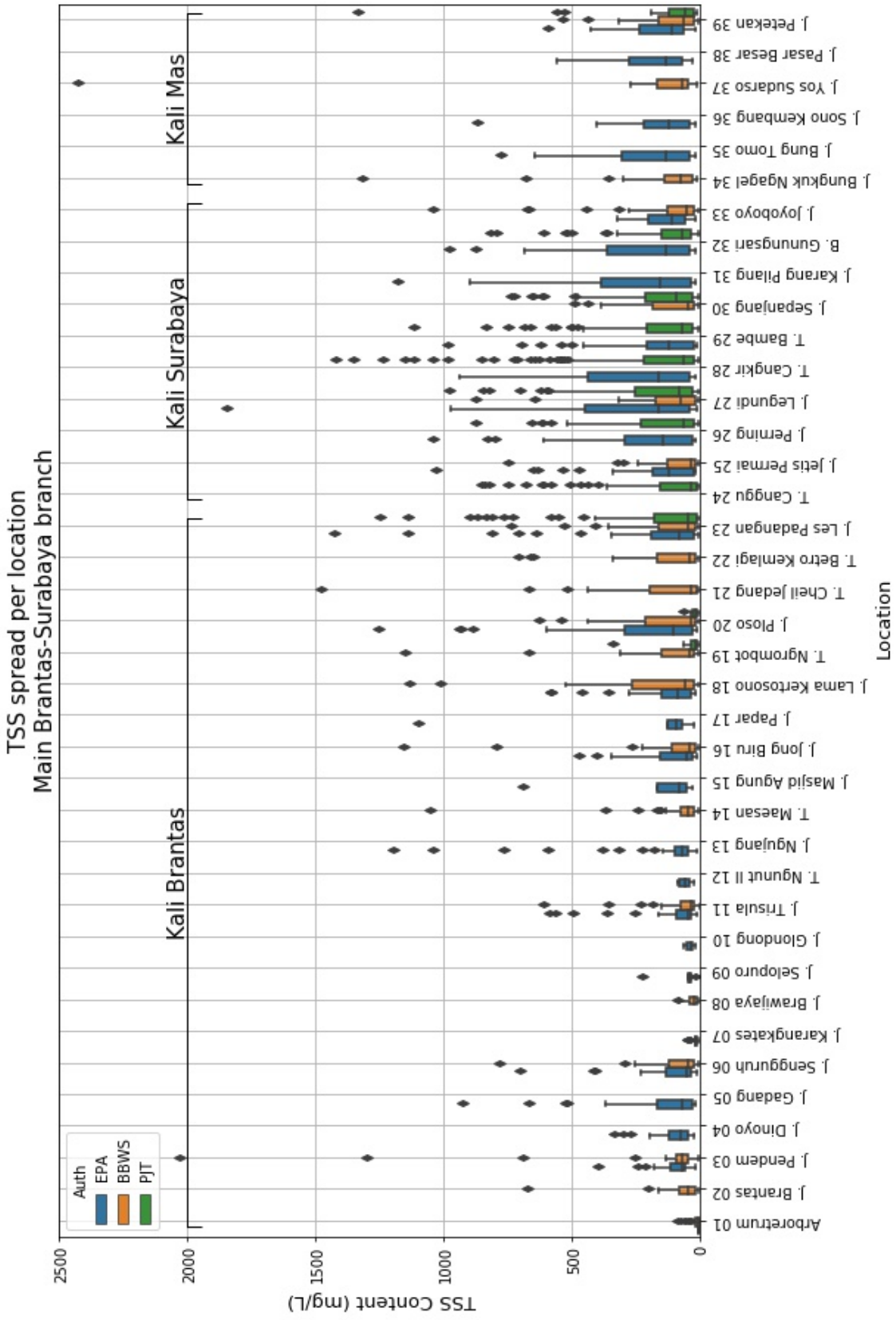

Figure B.13: TSS concentrations along the main Brantas stretch.

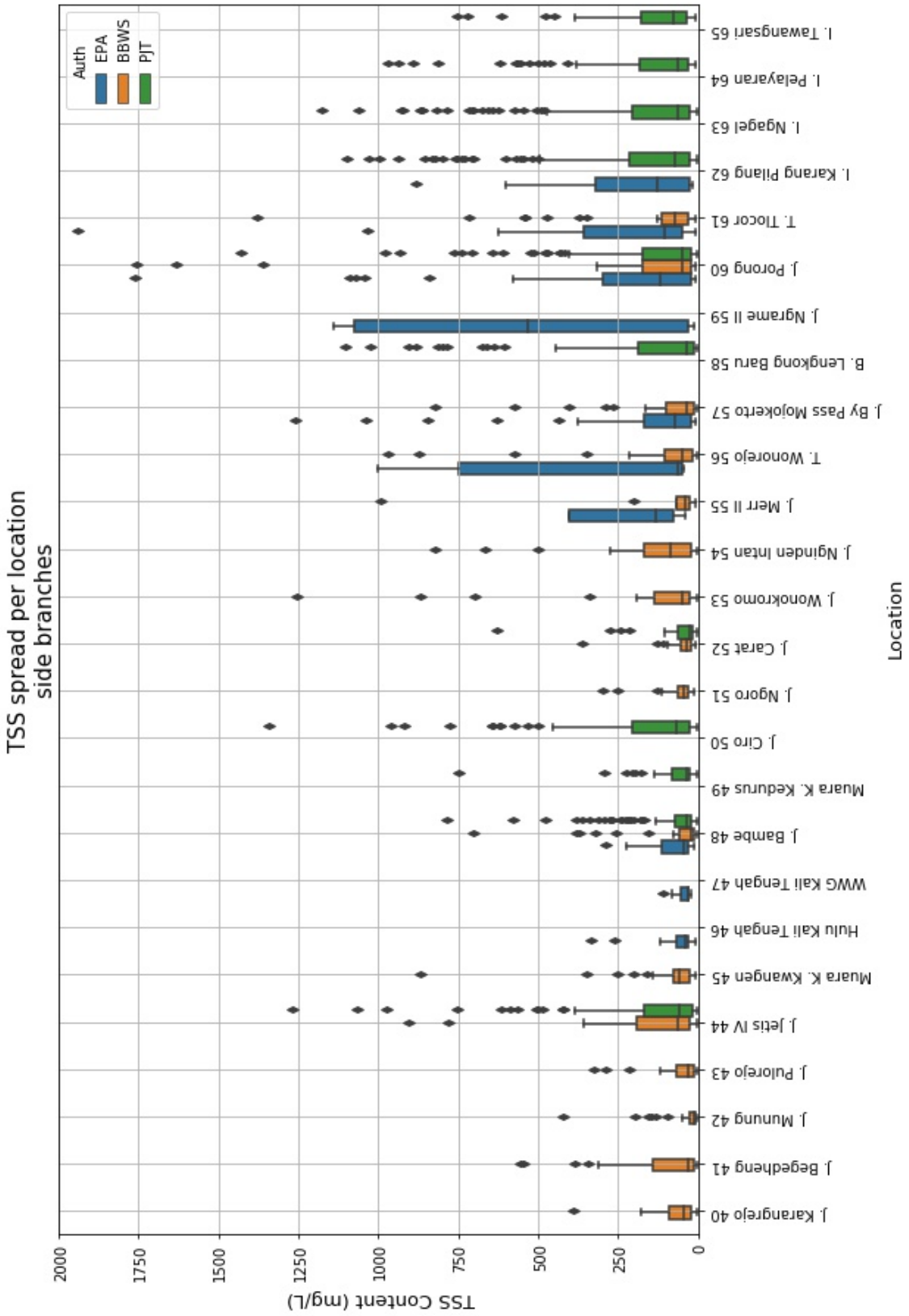

Figure B.14: TSS concentrations in the branches of the Brantas.

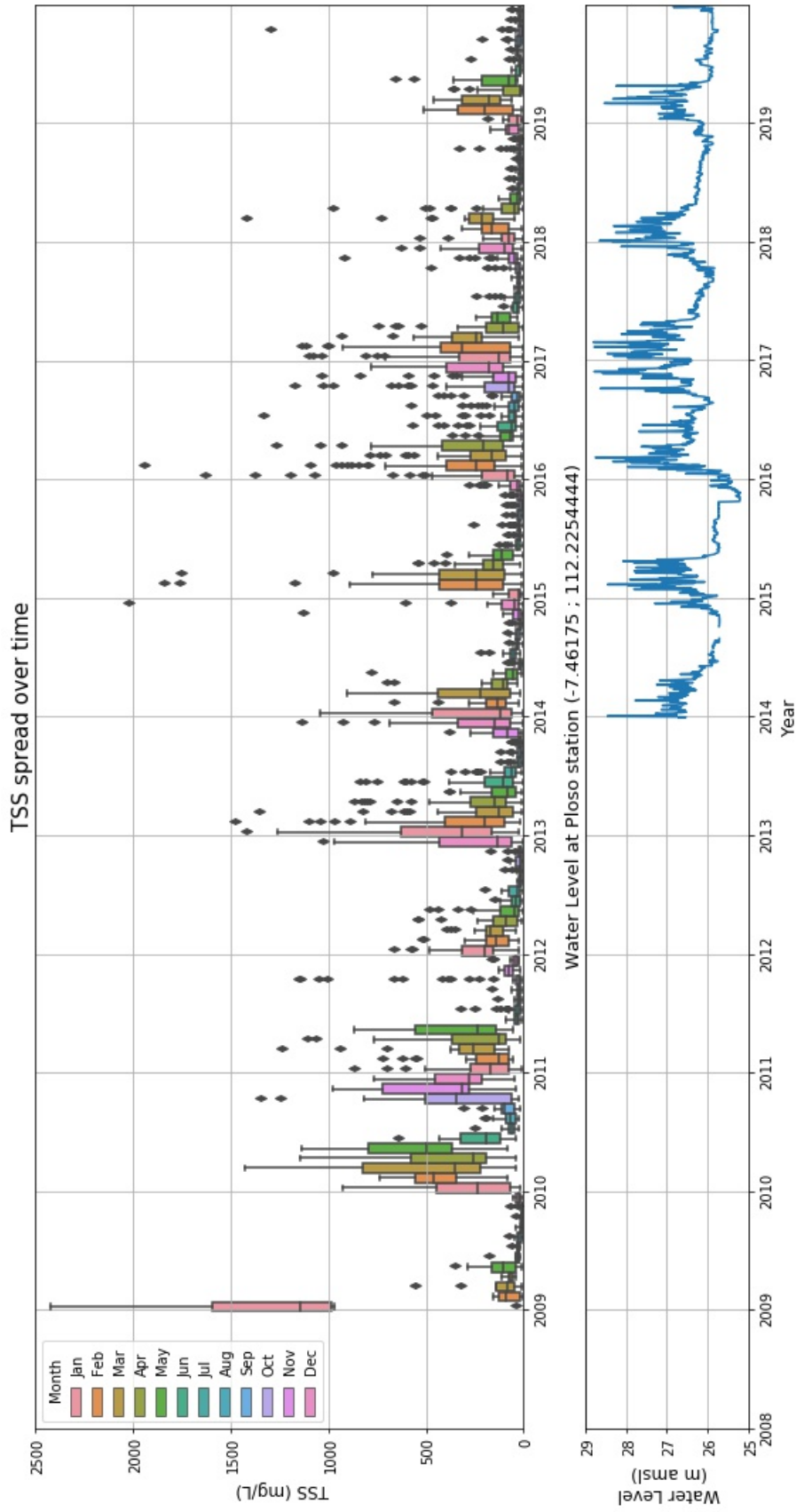

Figure B.15: TSS concentrations in the Brantas over time.

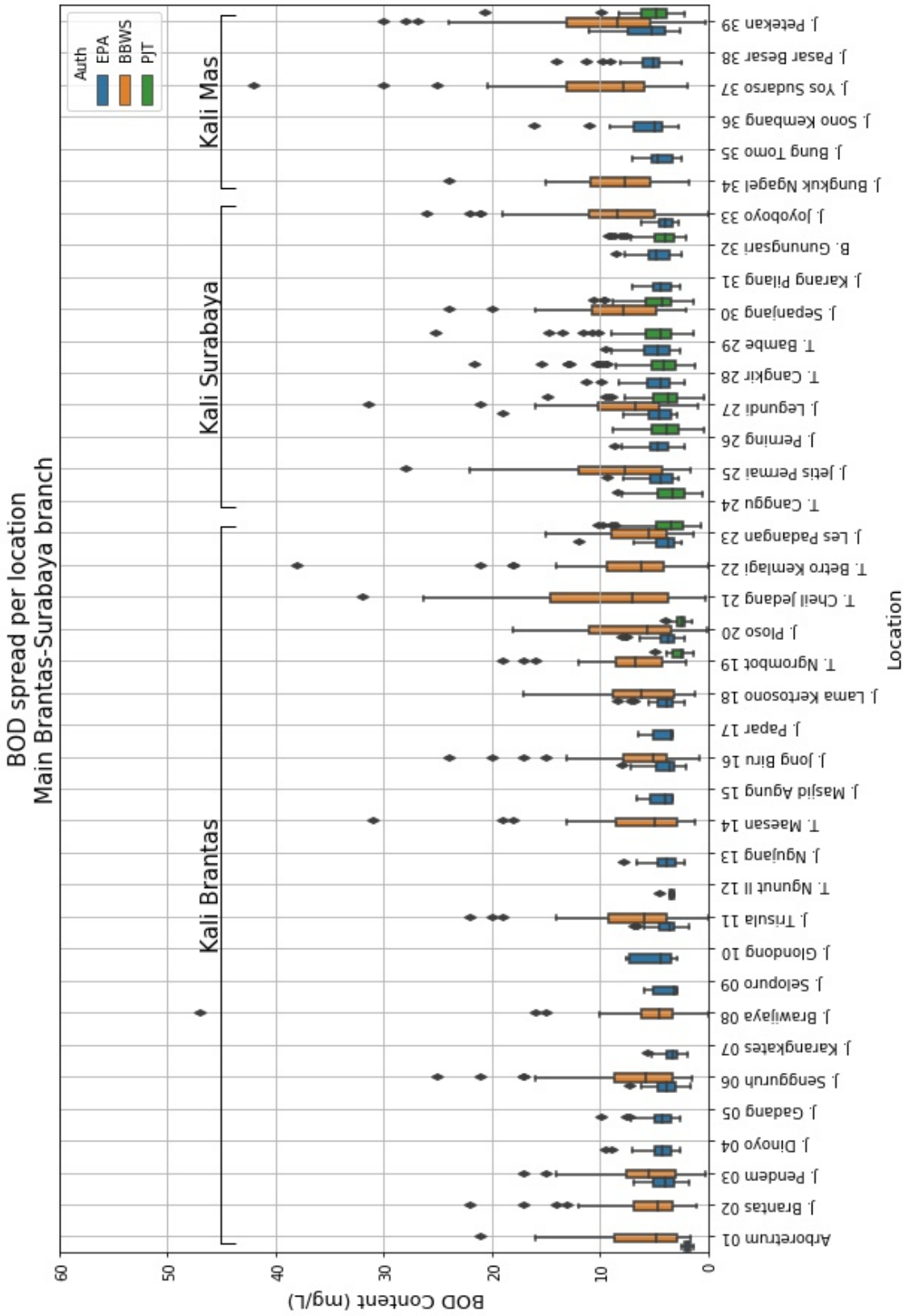

Figure B.16: BOD concentrations along the main Brantas stretch.

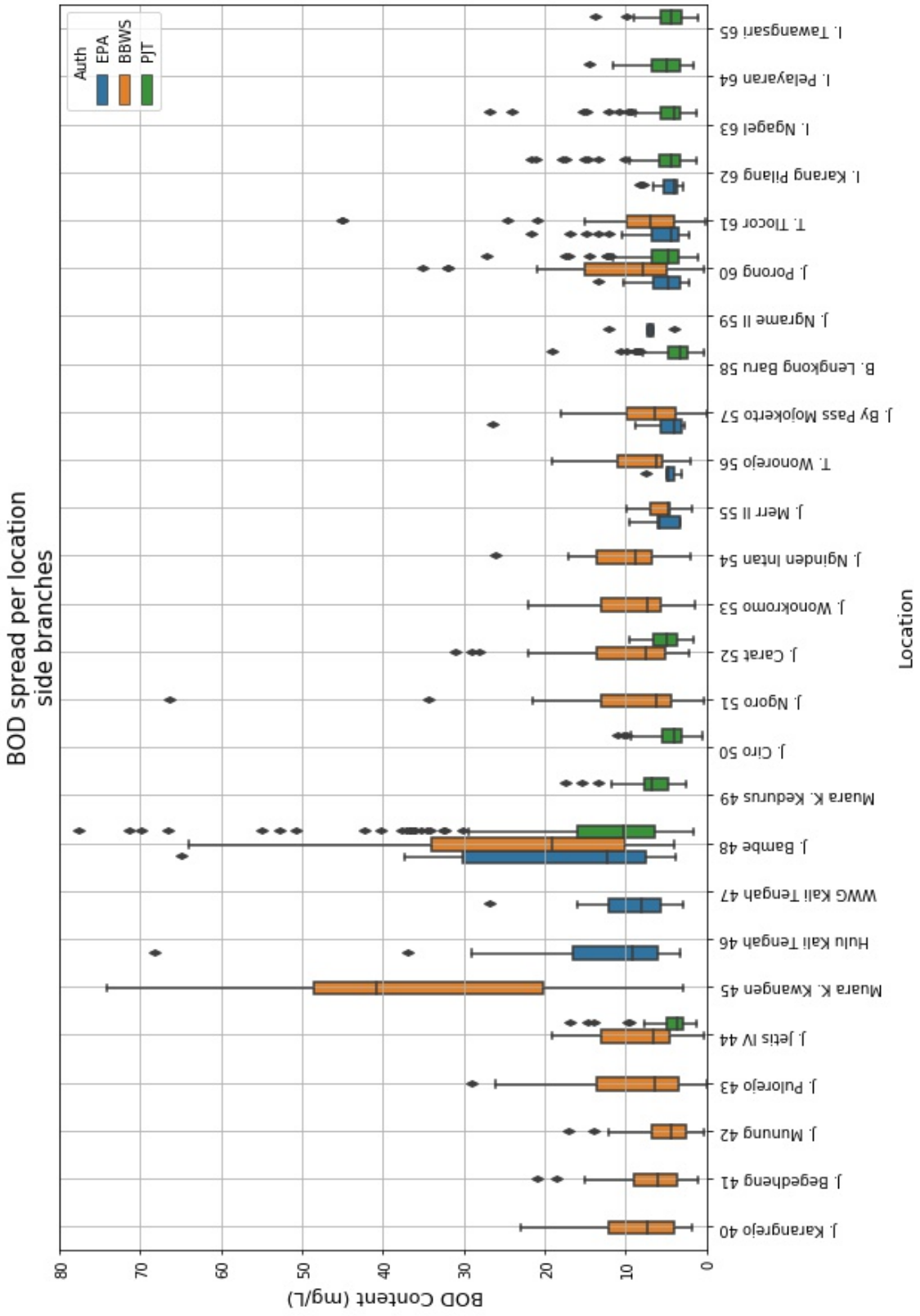

Figure B.17: BOD concentrations in the branches of the Brantas.

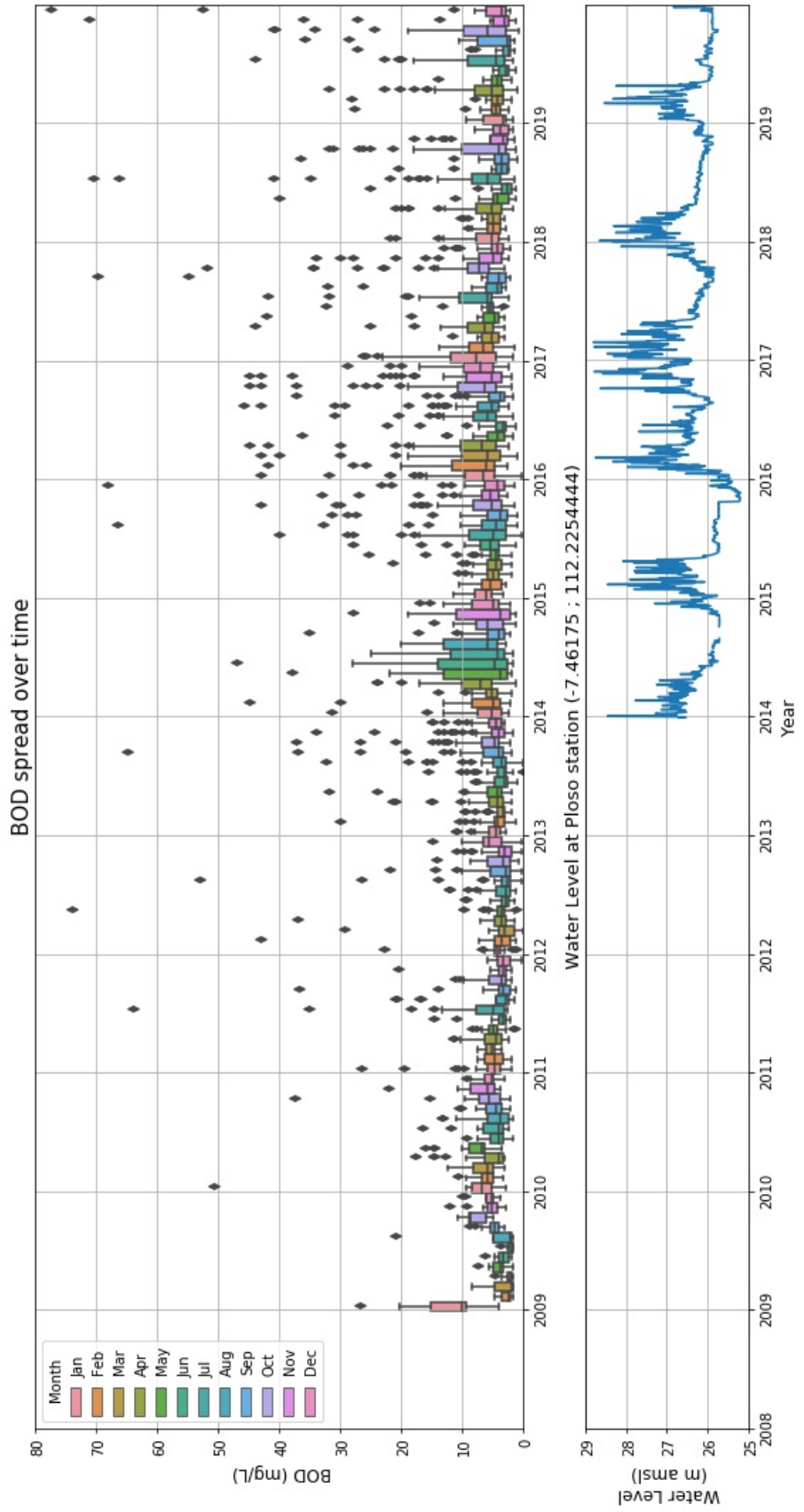

Figure B.18: BOD concentrations in the Brantas over time.

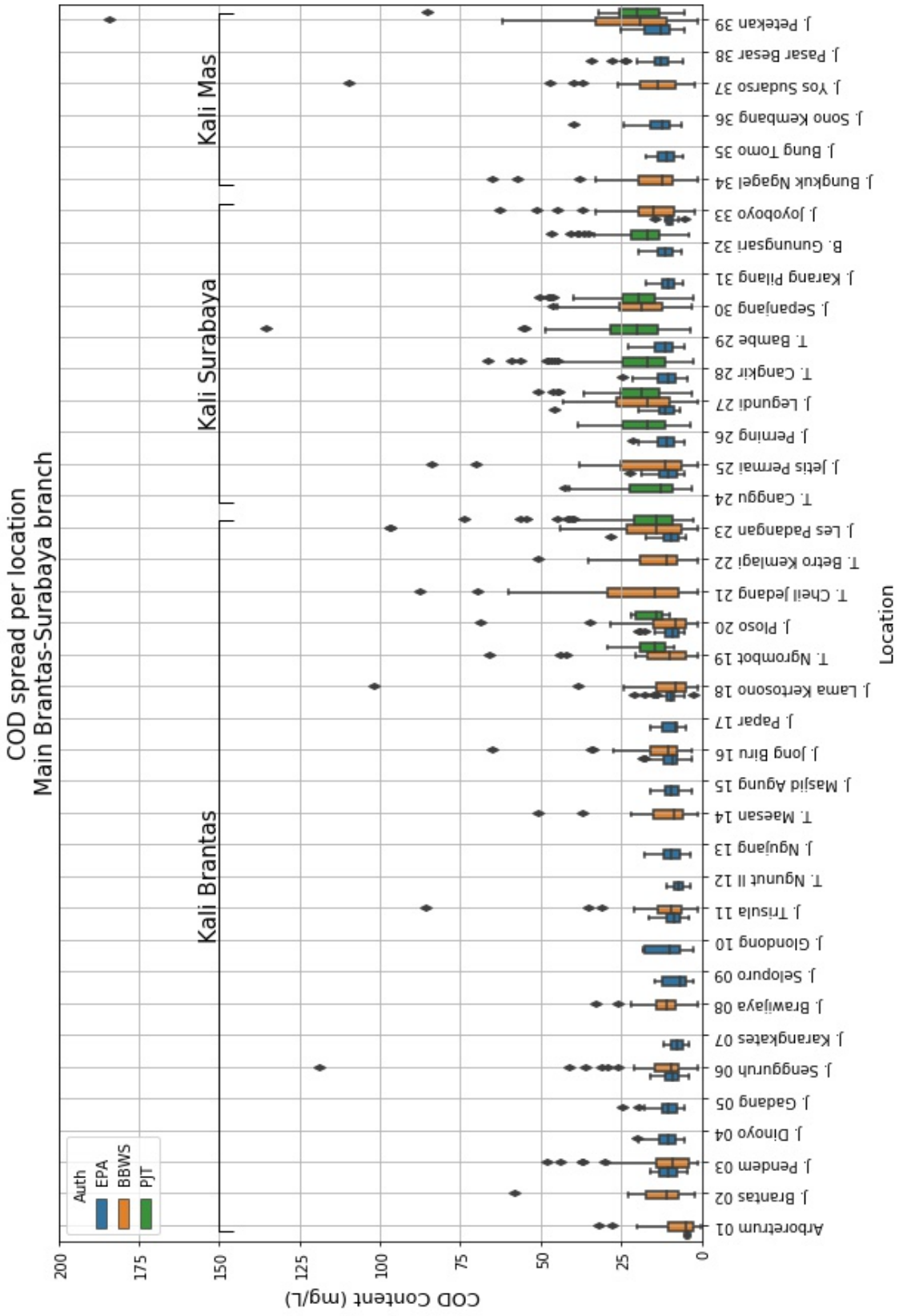

Figure B.19: COD concentrations along the main Brantas stretch.

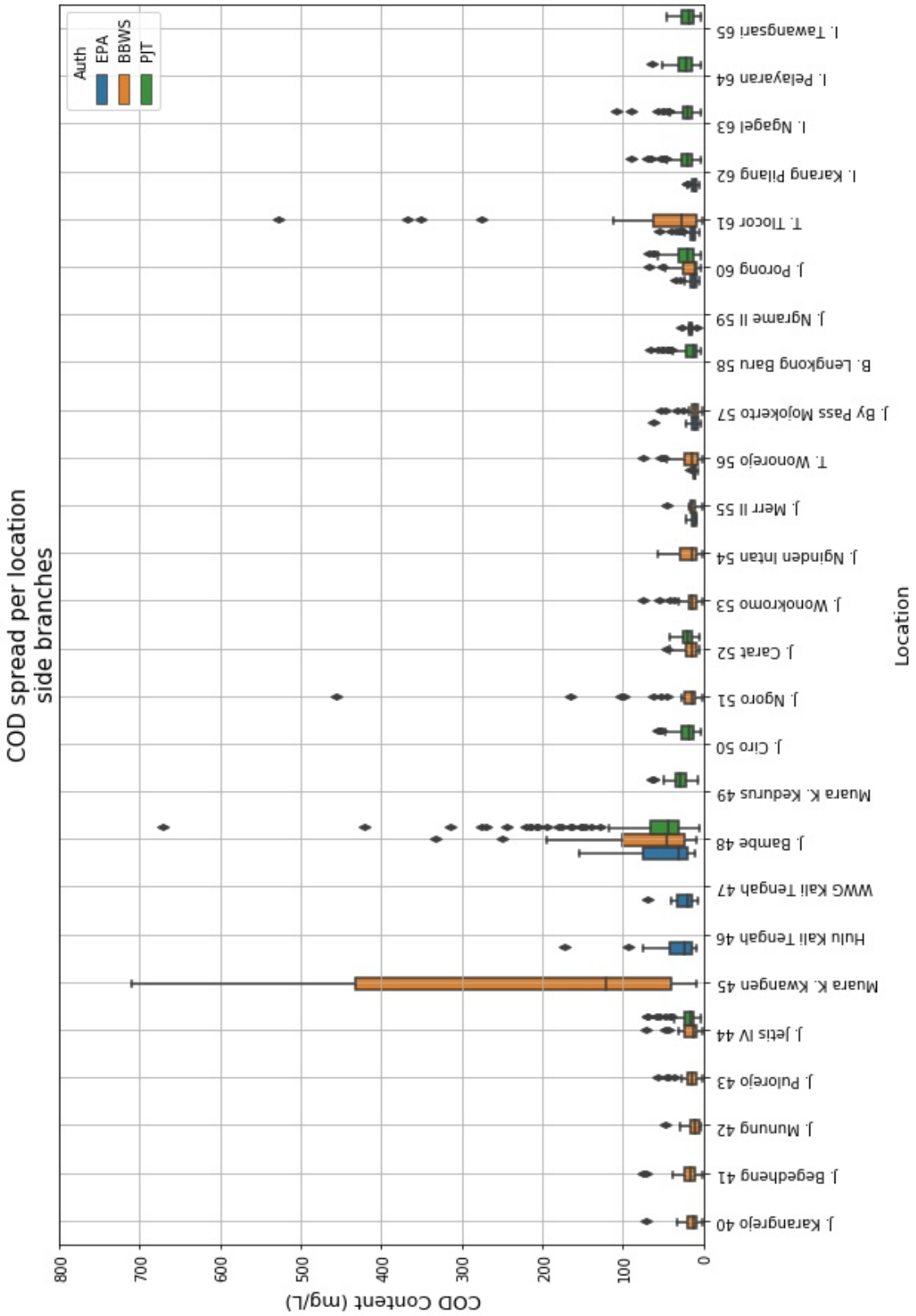

Figure B.20: COD concentrations in the branches of the Brantas.

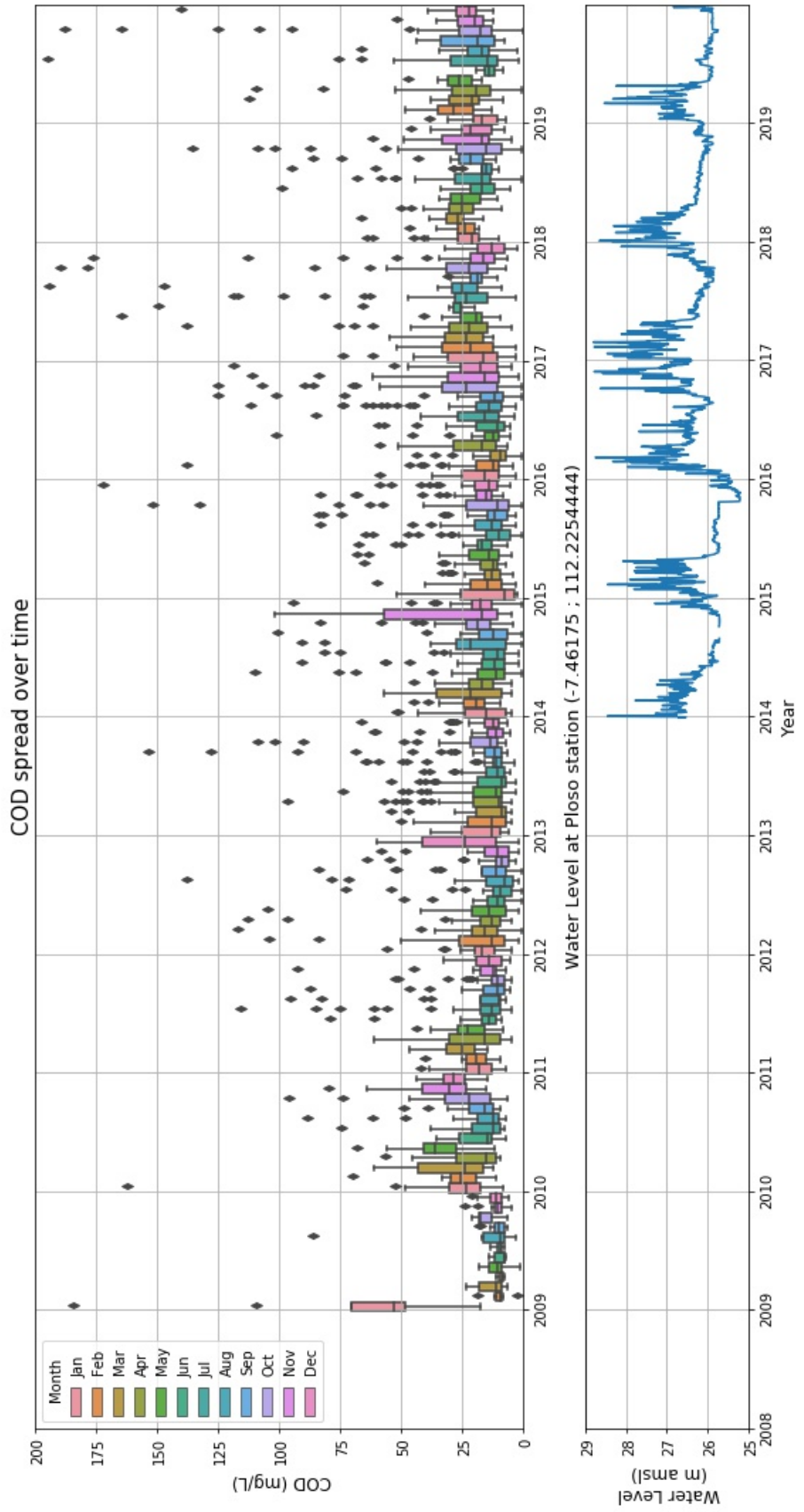

Figure B.21: COD concentrations in the Brantas over time.

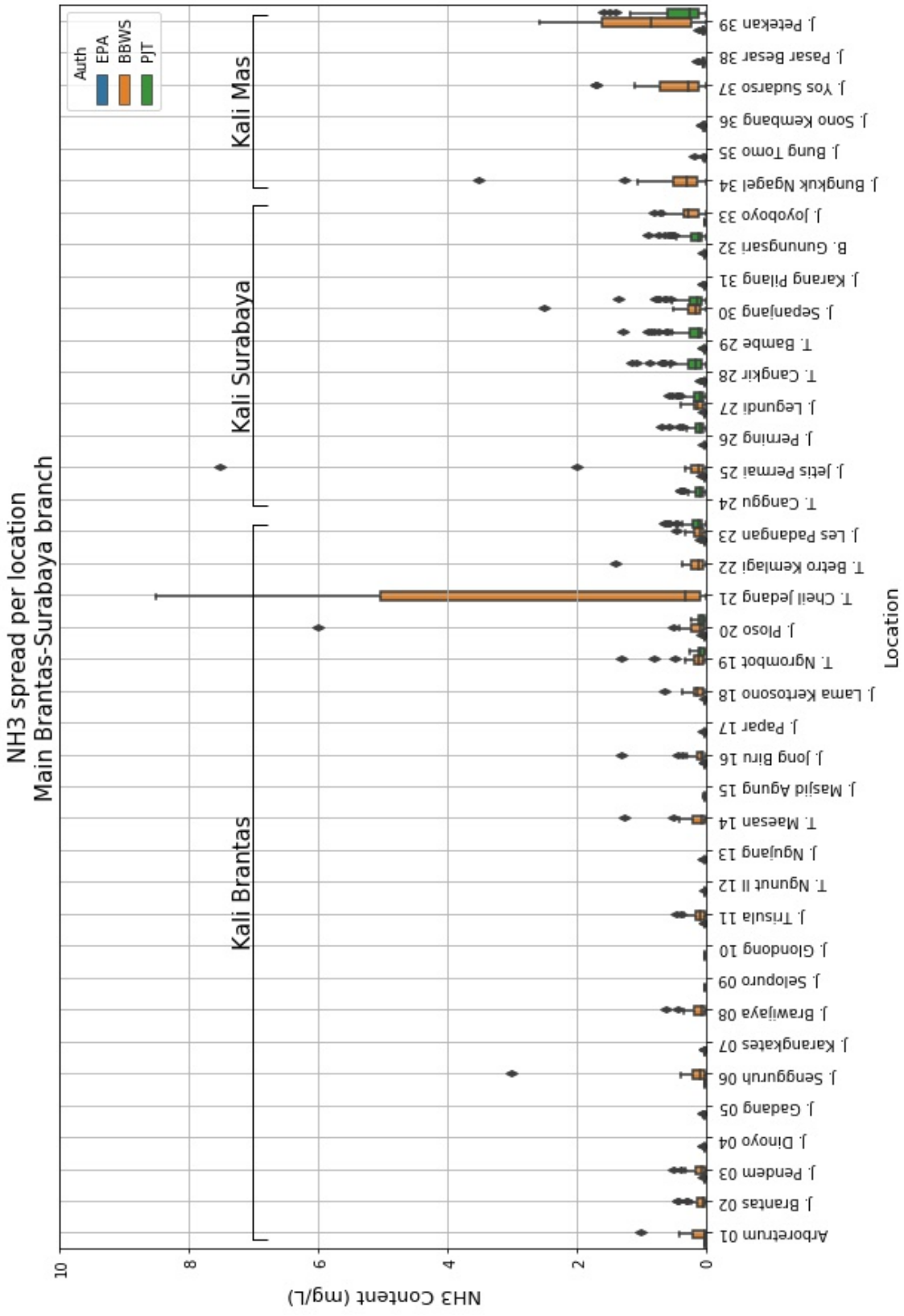

Figure B.22: NH_3 concentrations along the main Brantas stretch.

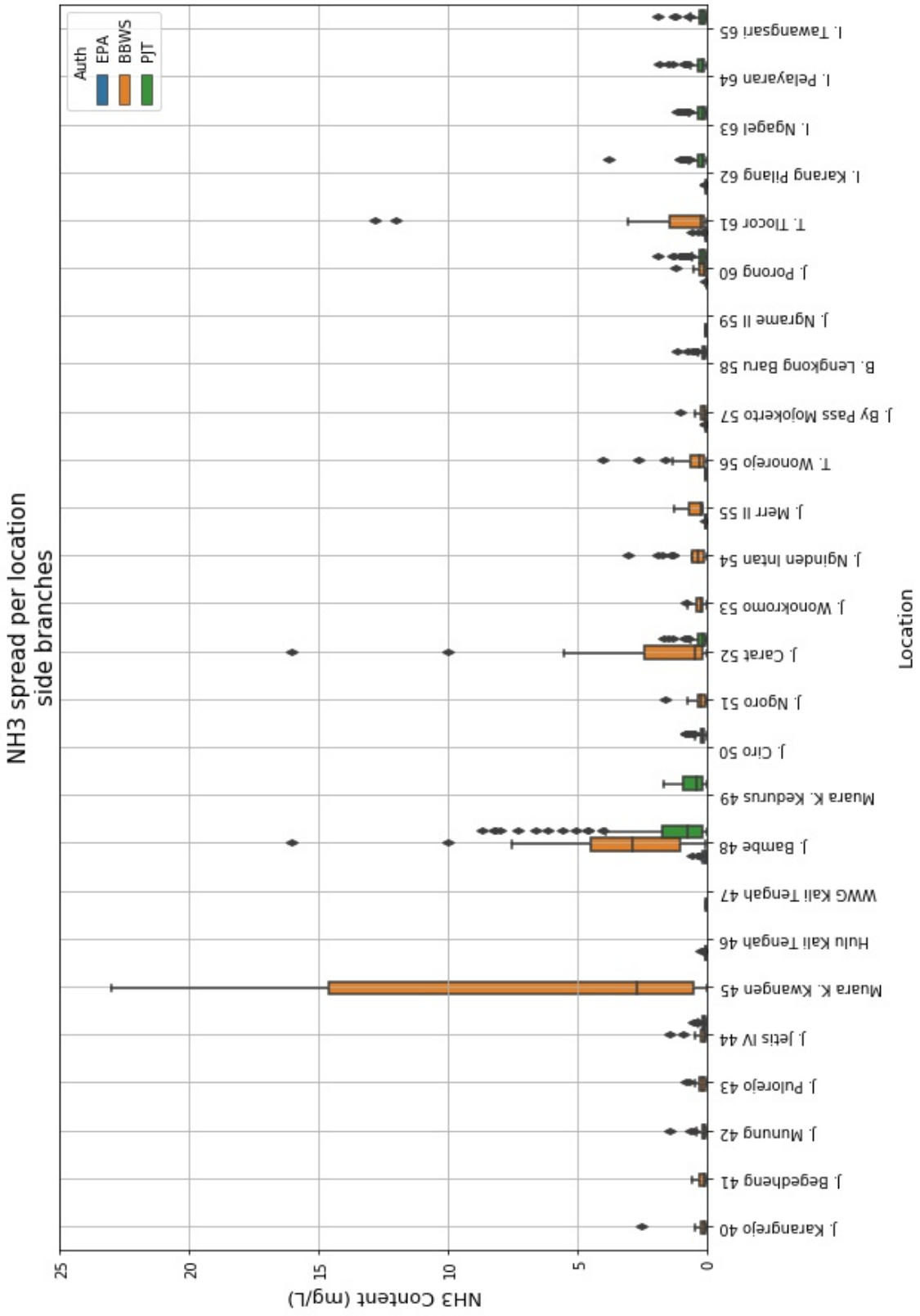

Figure B.23: NH_3 concentrations in the branches of the Brantas.

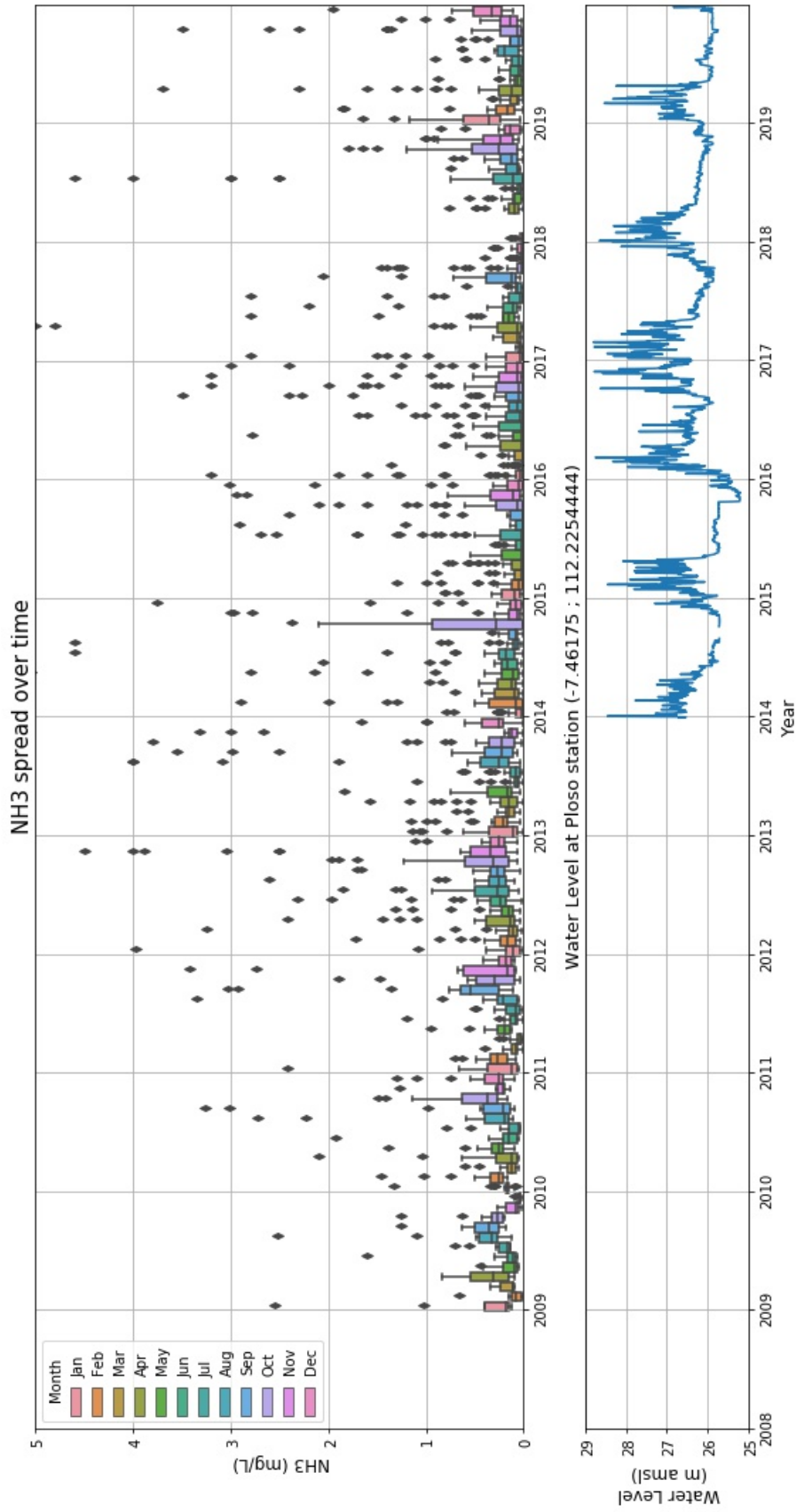

Figure B.24: NH_3 concentrations in the Brantas over time.

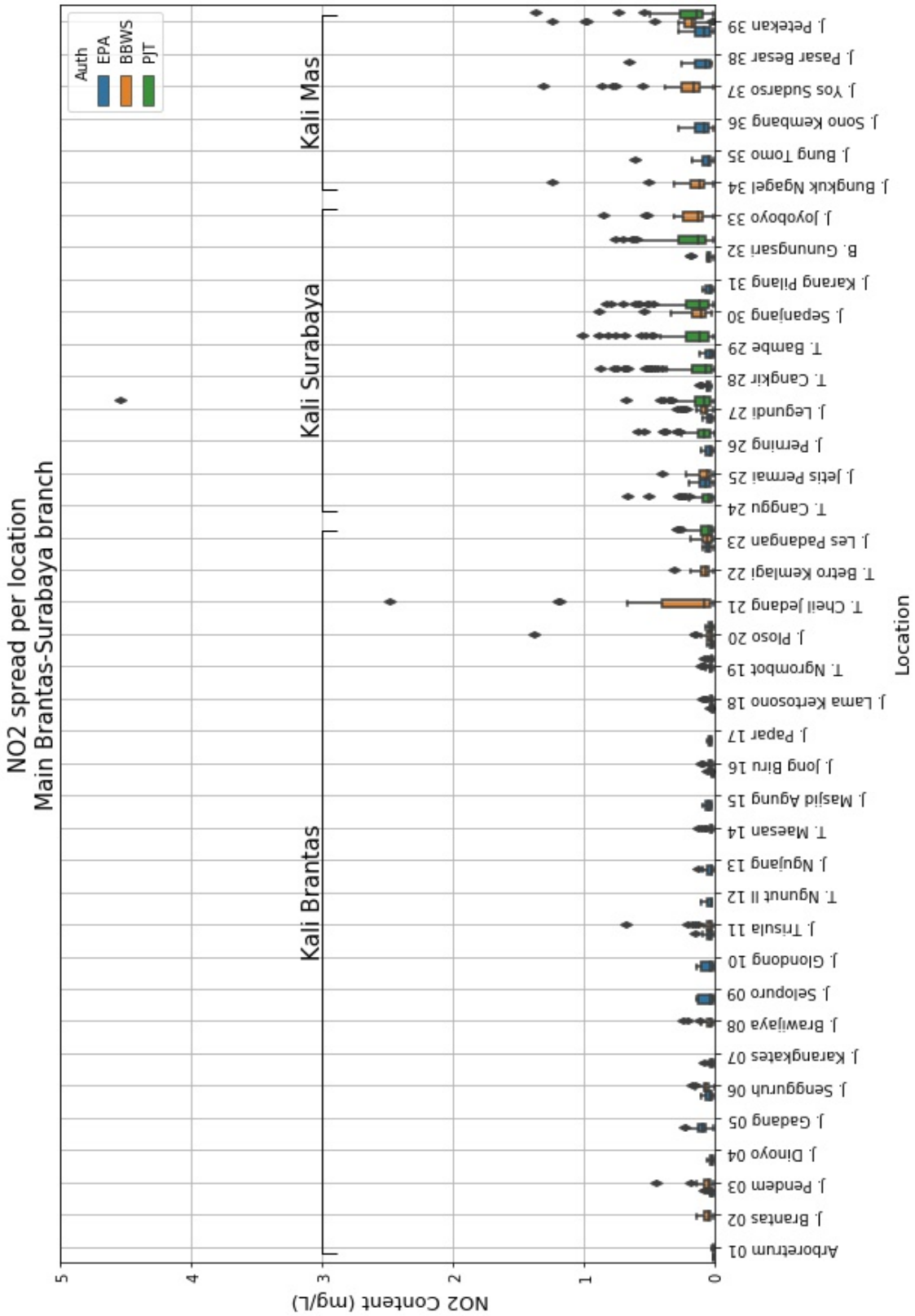

Figure B.25: NO_2 concentrations along the main Brantas stretch.

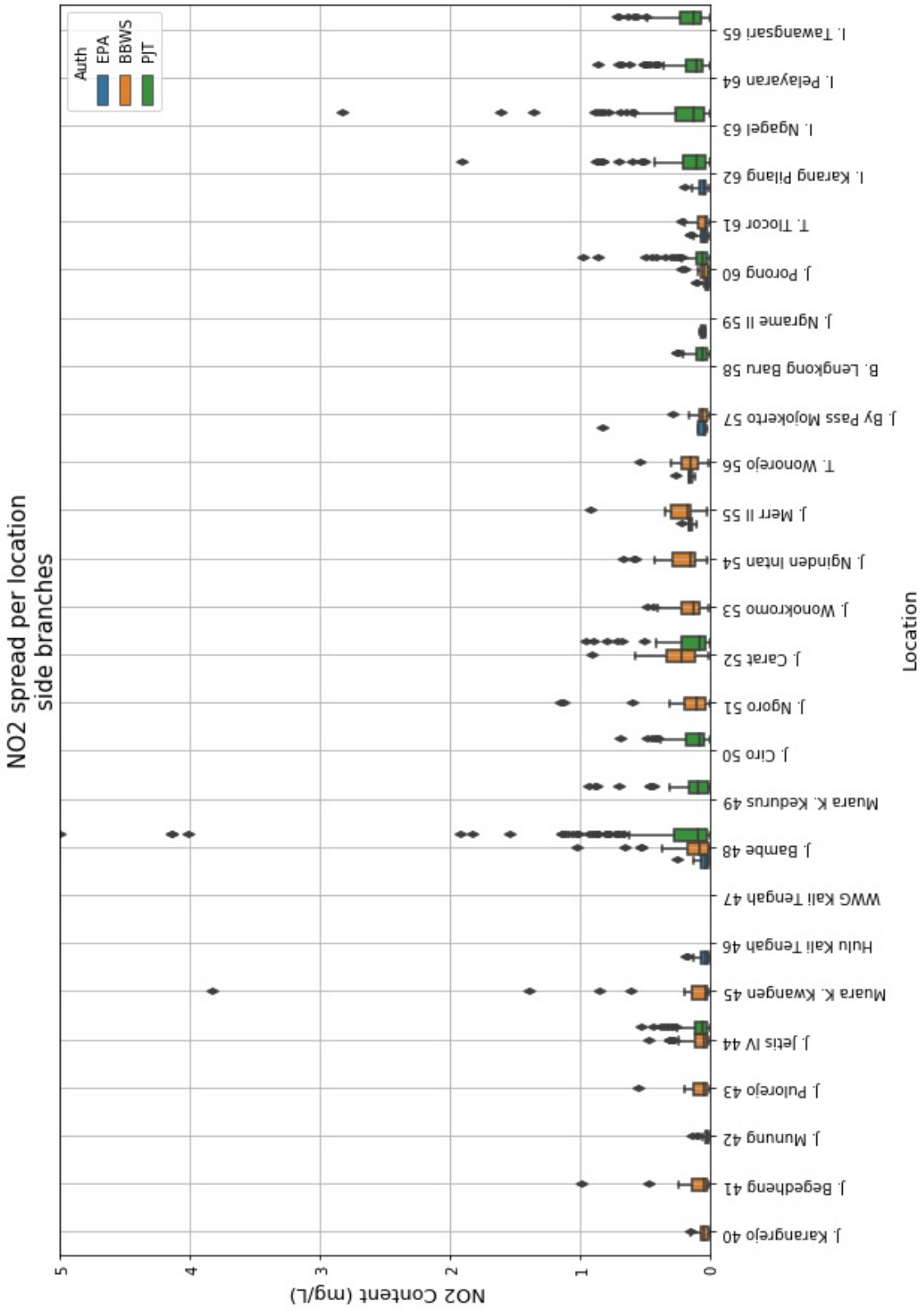

Figure B.26: NO_2 concentrations in the branches of the Brantas.

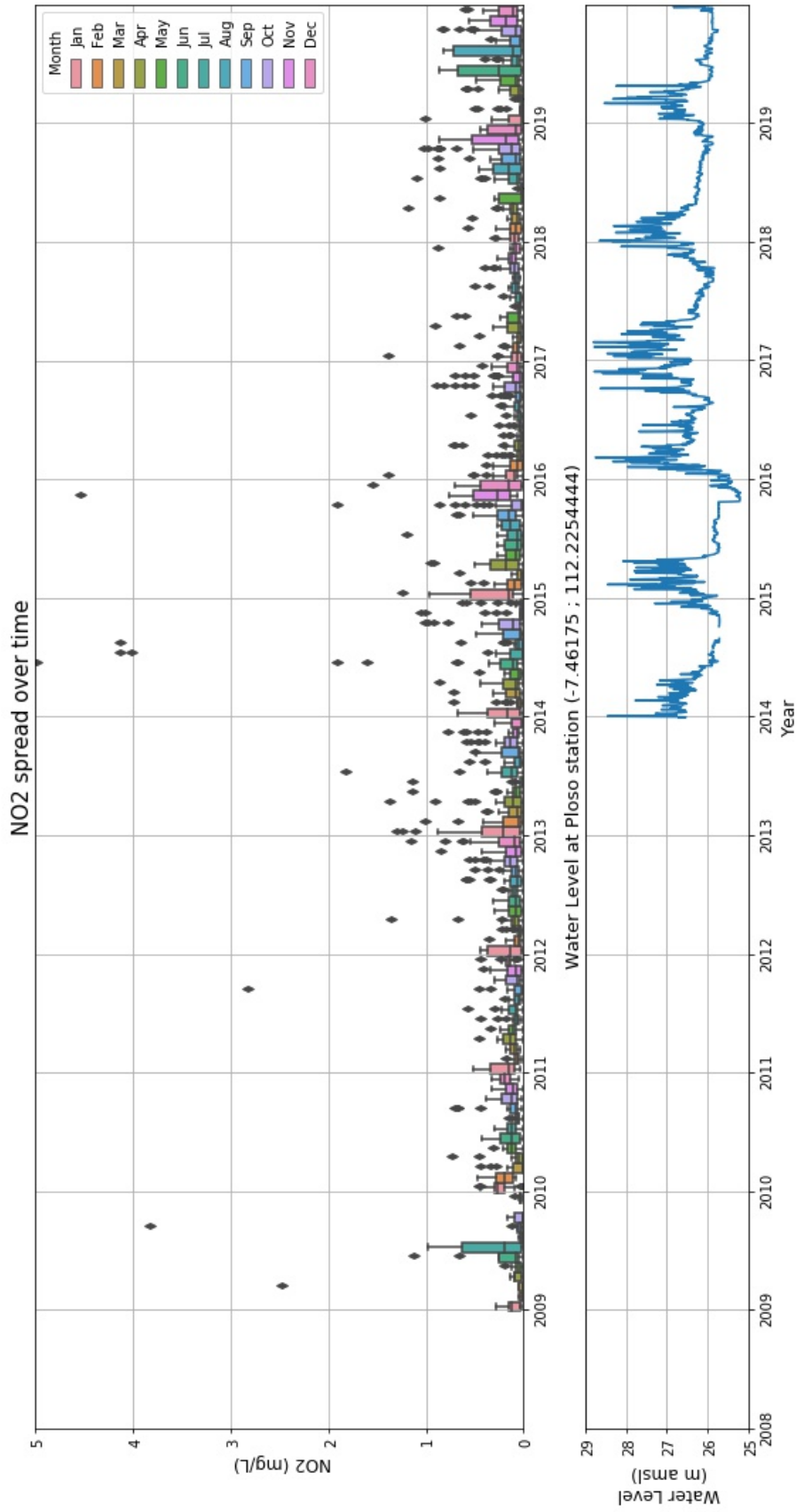

Figure B.27: NO_2 concentrations in the Brantas over time.

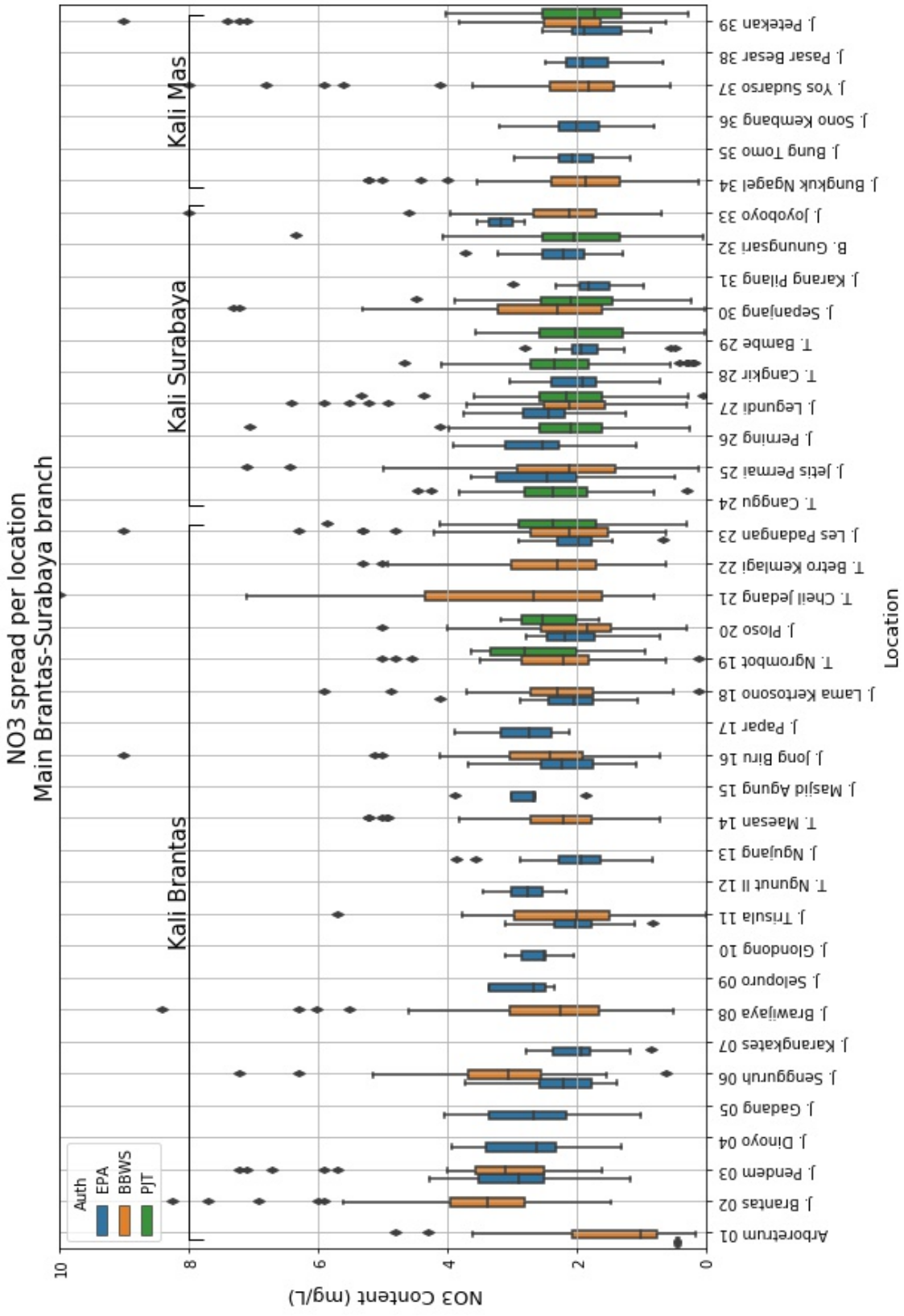

Figure B.28: NO_3 concentrations along the main Brantas stretch.

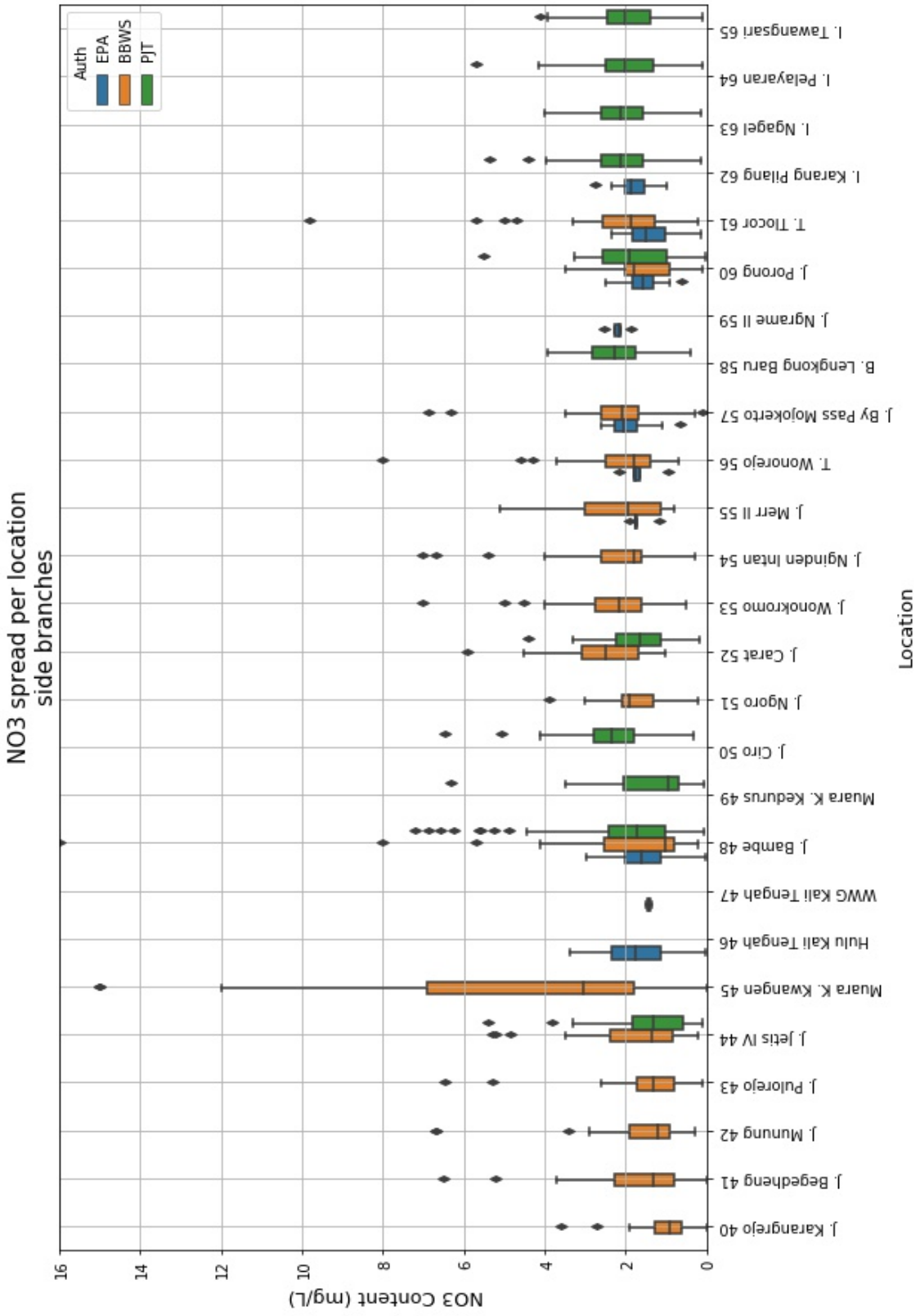

Figure B.29: NO_3 concentrations in the branches of the Brantas.

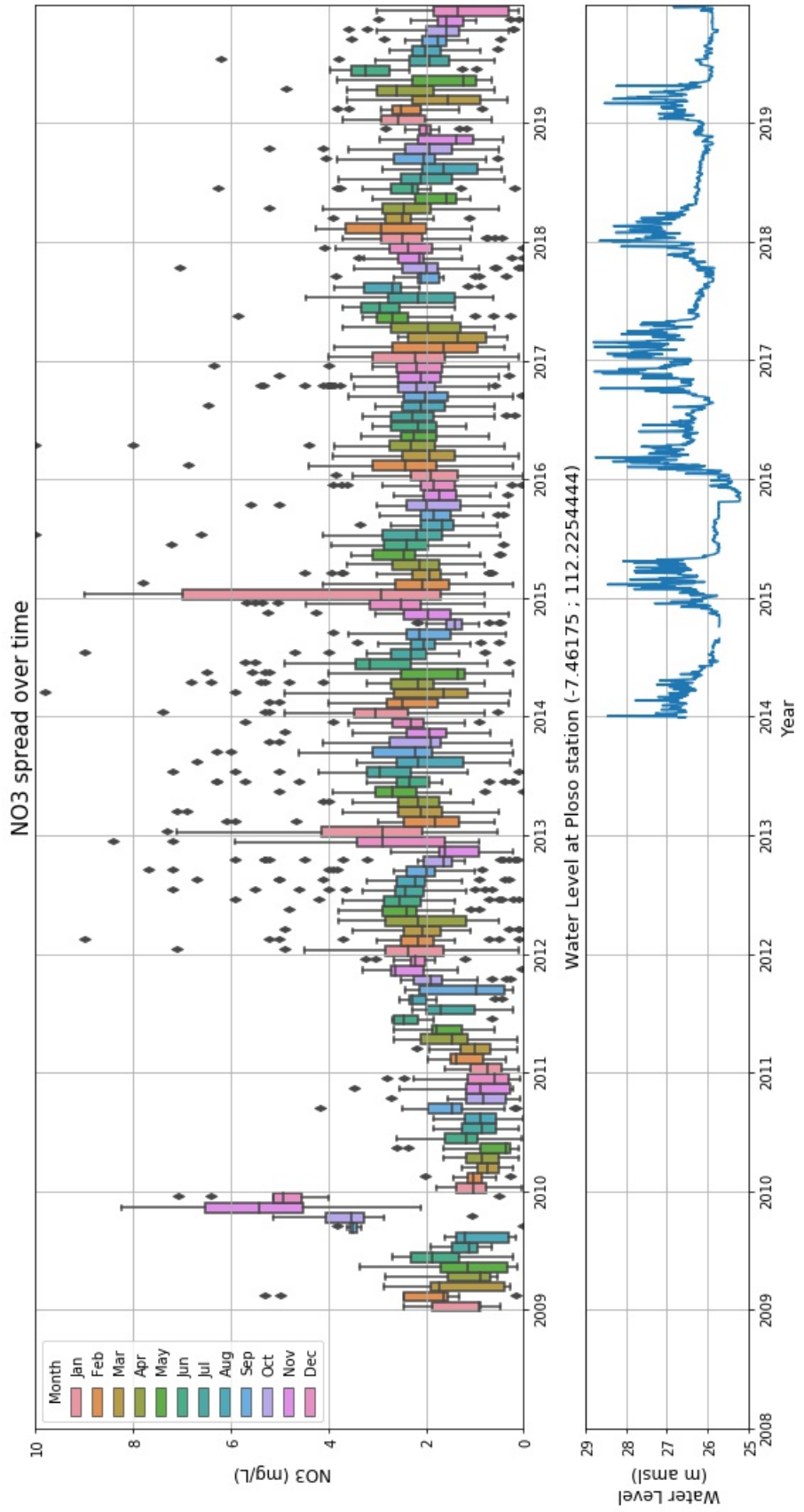

Figure B.30: NO_3 concentrations in the Brantas over time.

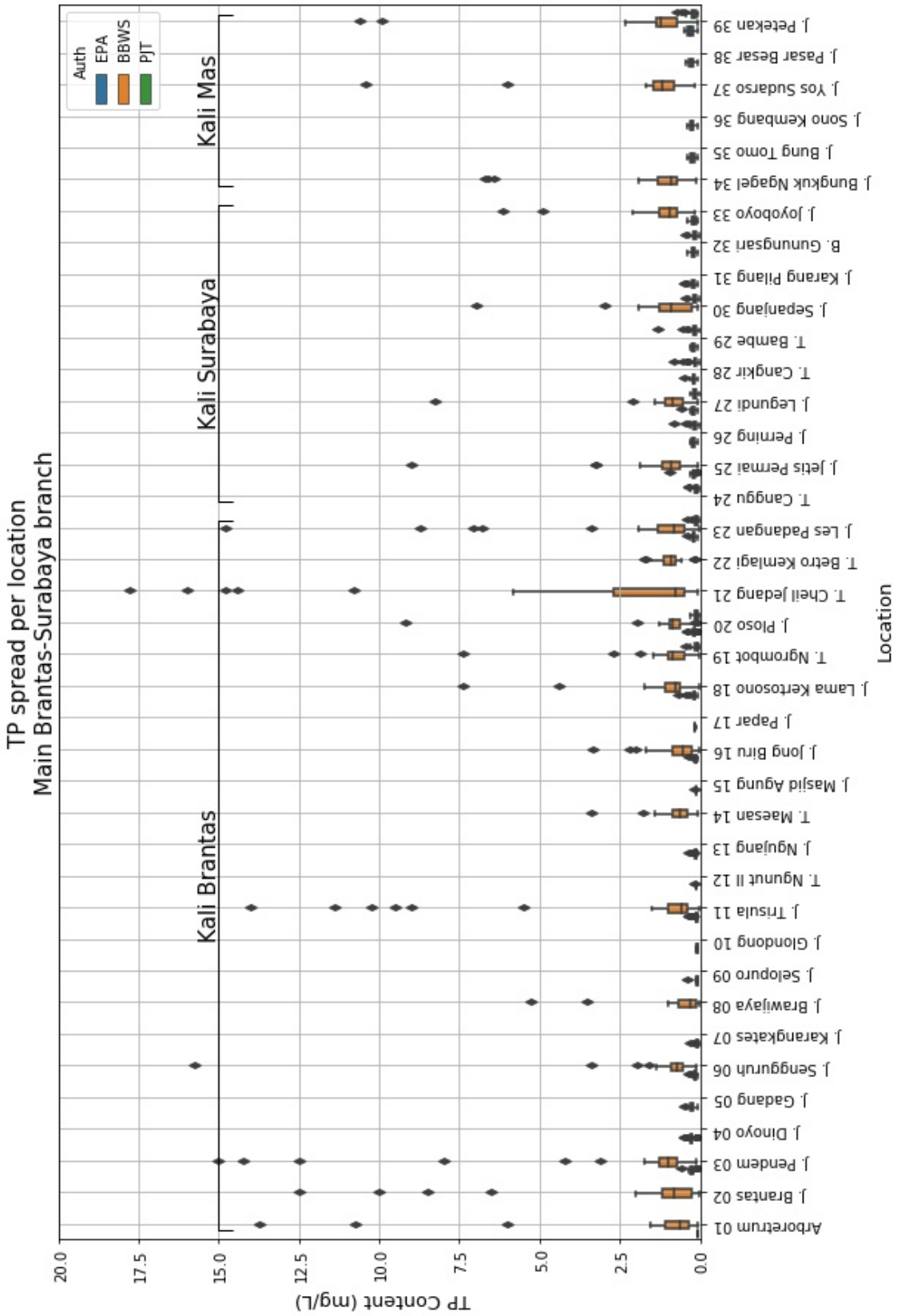

Figure B.31: Phosphate concentrations along the main Brantas stretch.

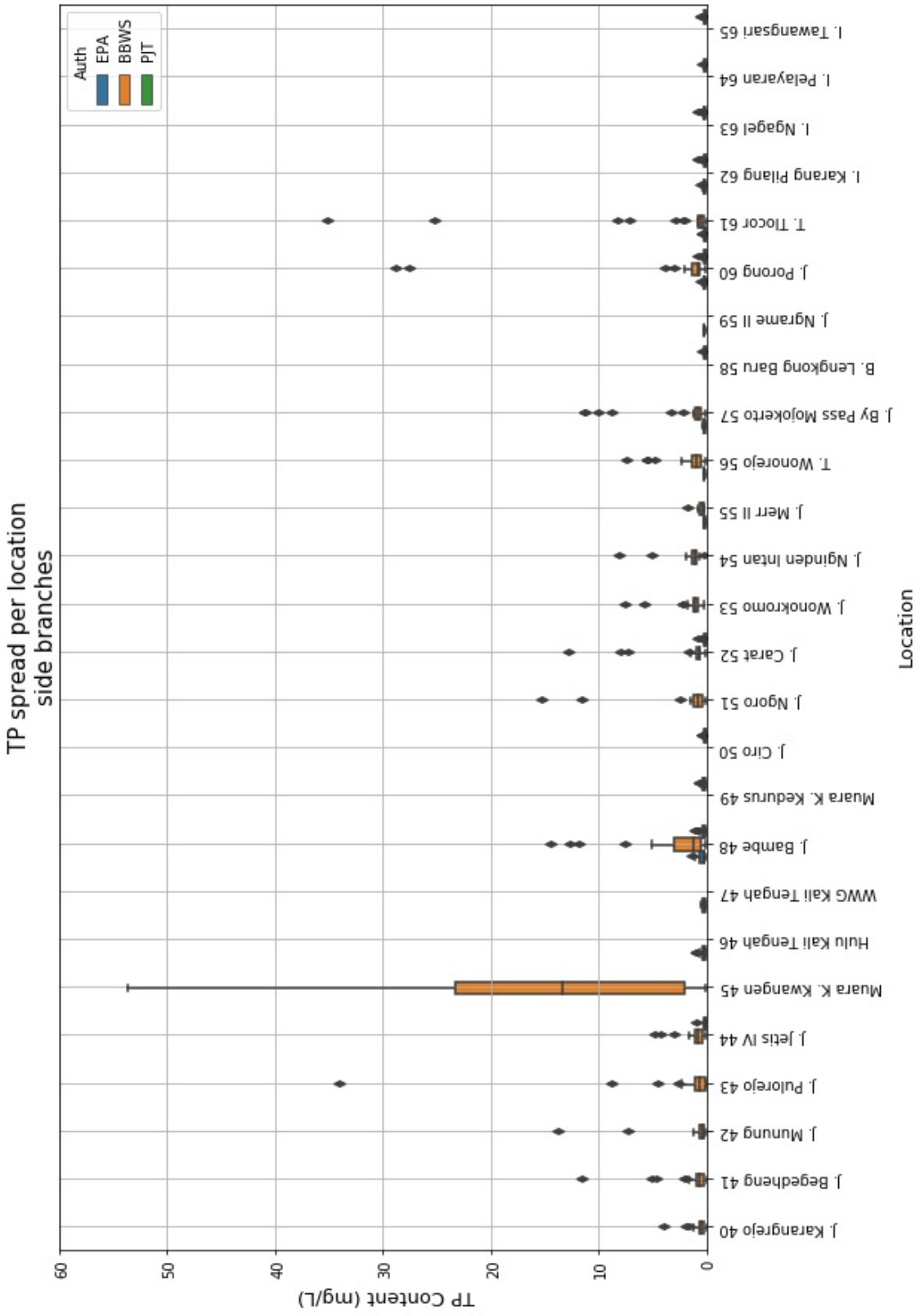

Figure B.32: Phosphate concentrations in the branches of the Brantas.

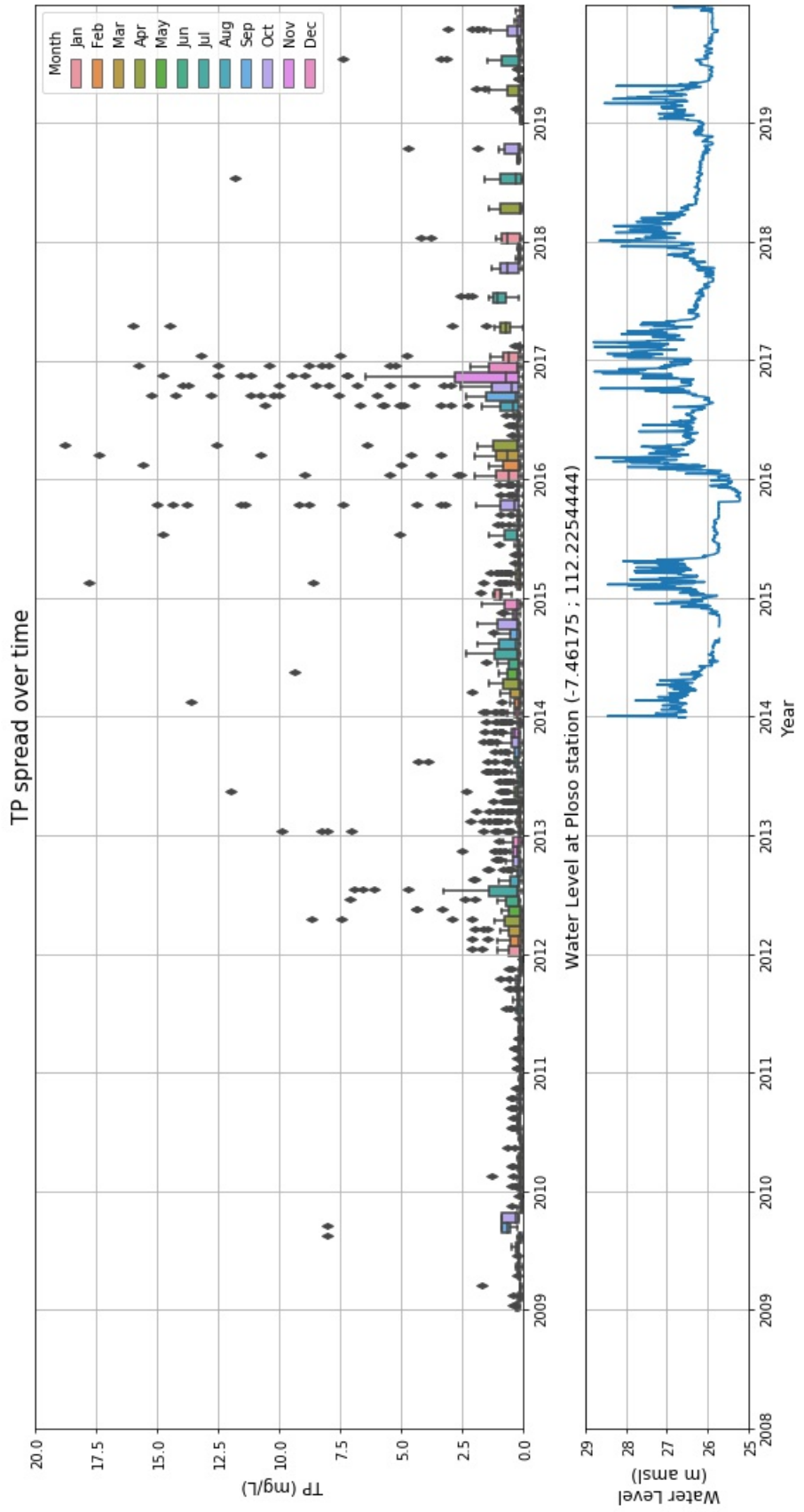

Figure B.33: Phosphate concentrations in the Brantas over time.

C LOGARITHMIC PCA RESULTS

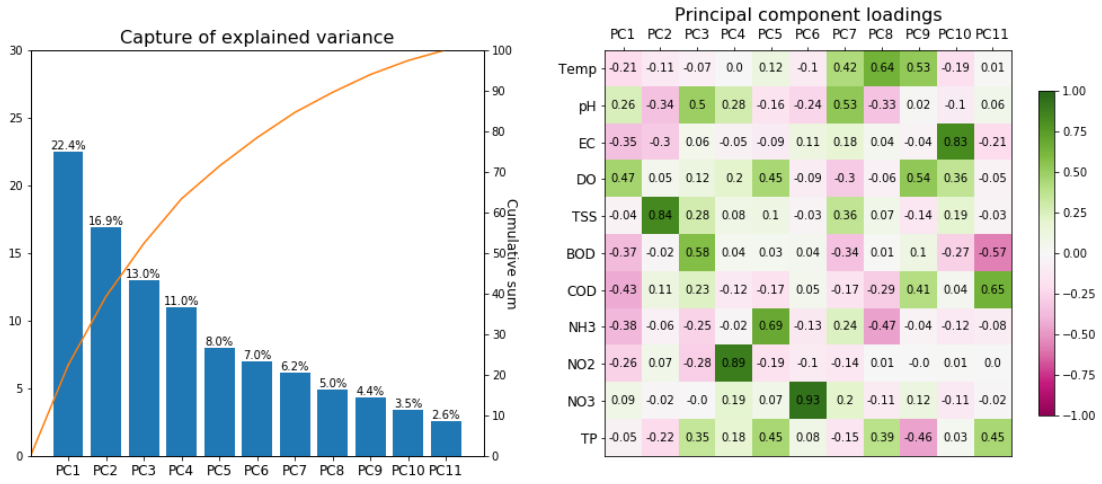

Figure C.1: PCA on dataset with logarithmic scaling, on the left is the capture of explained variance by principal component and on the right is the composition of said principal components' loadings.

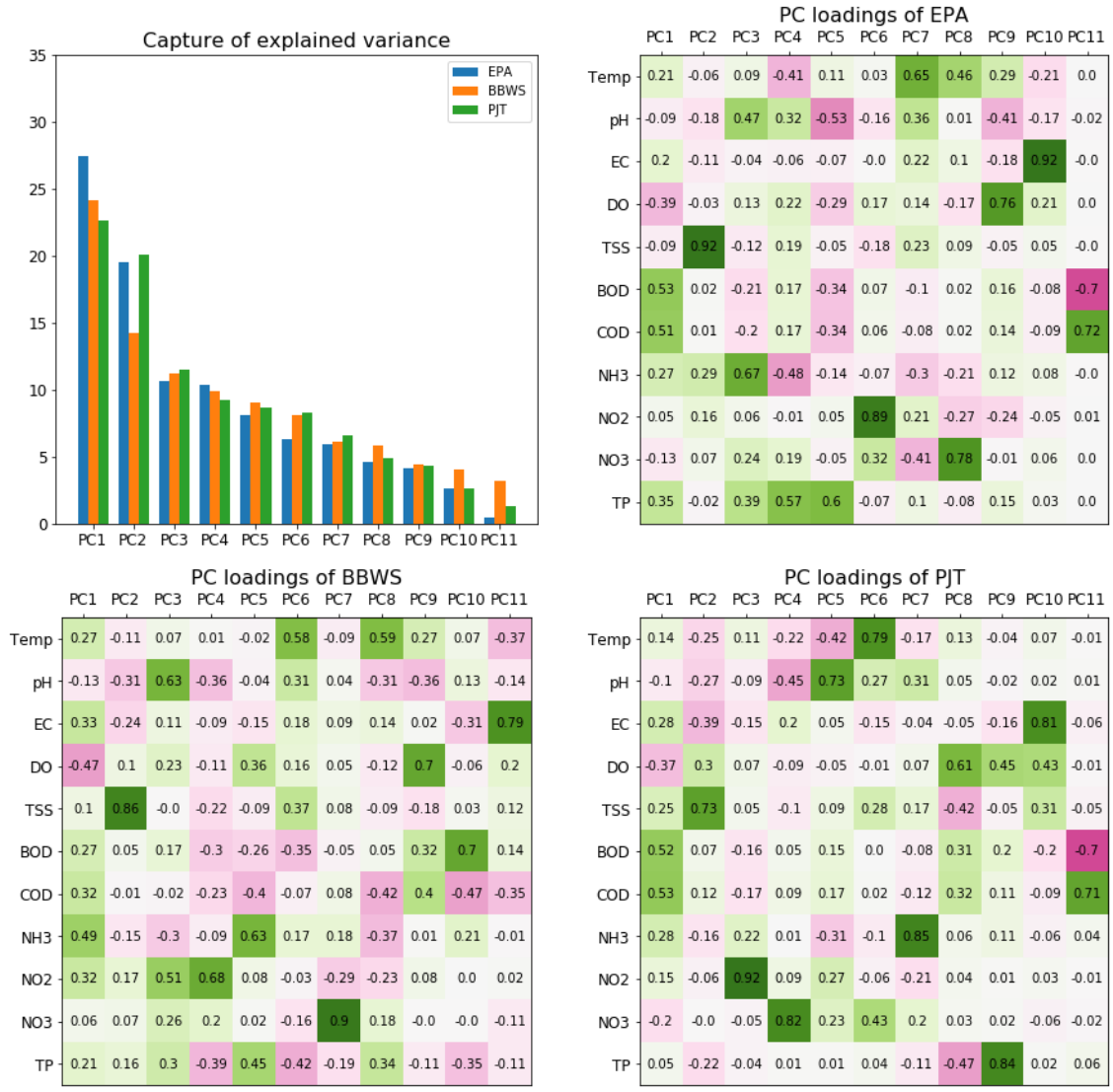

Figure C.2: PCA of separate authorities with logarithmic scaling. The top left graph depicts the capture of explained variance of the three different PCA's. The other 3 graphs depict the PCA loadings matrix of the specific authority.

D | PYTHON SCRIPTS

List of files with scripts used to produce results used in the thesis:

- Crossplots.ipynb
- Data_compile.ipynb
- Data_overview.ipynb
- Import_BBWS.ipynb
- Import_EPA.ipynb
- Import_PJT.ipynb
- Location_construct.ipynb
- PCA_AllData.ipynb
- PCA_byAuthority.ipynb
- PCA_MethodTesting.ipynb
- Statistics.ipynb
- Trajectory_plots.ipynb
- PCA.py

E DATA FILES

Datafiles will not be uploaded with the final thesis. The following is a list of files created:

- Composite_dataset.csv
- WL_PJT.csv
- BBWS/Actual_locations.csv
- Composite_BBWS.csv
- EPA/Actual_locations.csv
- Composite_EPA.csv
- PJT/Actual_locations.csv
- Composite_PJT.csv

And these are the files that were used to create these:

- AWLR Gadang - Malang 2014-2019.xlsx
- AWLR Jeli - dekat Kediri 2014-2019.xlsx
- AWLR Kediri 2014-2019.xlsx
- AWLR Perning - sesudah Mojokerto 2014-2019.xlsx
- AWLR Ploso - sebelum Mojokerto 2014-2019.xlsx
- AWLR Porong 2014-2019.xlsx
- BBWS_2009.xls
- BBWS_2010.xls
- BBWS_2011.xls
- BBWS_2012.xlsx
- BBWS_2013.xlsx
- BBWS_2014.xlsx
- BBWS_2015.xlsx
- BBWS_2016.xlsx
- BBWS_2017.xlsx
- BBWS_2018.xlsx
- BBWS_2019.xlsx

- BRANTAS_WQ_2013.xlsx
- BRANTAS_WQ_2014.xlsx
- BRANTAS_WQ_2016.xlsx
- BRANTAS_WQ_2017.xlsx
- Kualitas Air Brantas Hilir Sungai (2010).xlsx
- Kualitas Air Brantas Hilir Sungai (2011).xlsx
- Kualitas Air Brantas Hilir Sungai (2012).xlsx
- Kualitas Air Brantas Hilir Sungai (2013).xlsx
- Kualitas Air Brantas Hilir Sungai (2014).xlsx
- Kualitas Air Brantas Hilir Sungai (2015).xlsx
- Kualitas Air Brantas Hilir Sungai (2016).xlsx
- Kualitas Air Brantas Hilir Sungai (2017).xlsx
- Kualitas Air Brantas Hilir Sungai (2018).xlsx
- Kualitas Air Brantas Hilir Sungai (2019).xlsx

COLOPHON

This document was typeset using LaTeX. The document layout was generated using the arsclassica package by Lorenzo Pantieri, which is an adaption of the original classicthesis package from André Miede.