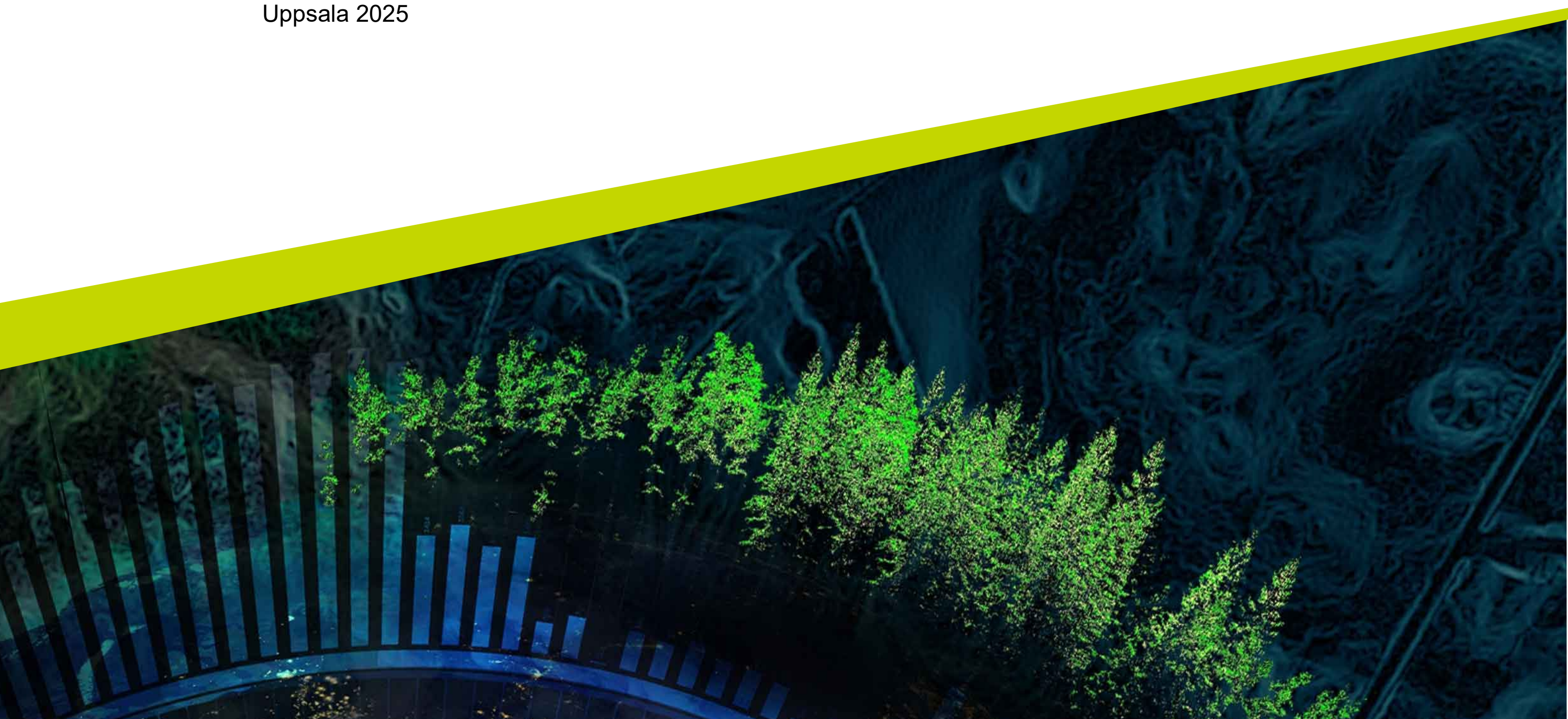
How do catchment characteristics of the National water monitoring system compare to catchment characteristics of Raw water sources in Sweden?

Carl Thurell

Självständigt arbete • 15 hp
Swedish University of Agricultural Sciences, SLU
BioMiljö
Uppsala 2025



How do catchment characteristics of the National water monitoring system compare to catchment characteristics of Raw water sources in Sweden?

Carl Thurell

Supervisor:	**Anna Lackner, SLU, Institutionen för vatten och miljö**
Examiner:	**Kevin Bishop, SLU, Institutionen för vatten och miljö**

Credits:	15p
Level:	G2E
Course title:	Självständigt arbete i Biologi
Course code:	EX0894
Programme/education:	BioMiljöprogrammet
Course coordinating dept:	Institutionen för vatten och miljö
Place of publication:	Uppsala
Year of publication:	2025

Keywords: Catchment characteristics, National water monitoring system, raw water sources, Trendsjöar, Omdrevssjöar

Swedish University of Agricultural Sciences

Department of Aquatic Sciences and Assessment

Abstract

The chemical composition and surrounding land use of bodies of water in a catchment are vital factors concerning how water within lakes can be used. Sweden has for a long time monitored these factors via the national water monitoring system, resulting in a large time series of data collected with the same method in a consistent way. These data are important, since it is used to draw overarching conclusions about the state of Swedish water. One of the use cases of this data would be drawing conclusions concerning the Swedish drinking water, as having a large overarching set of data would provide insights into the condition of our drinking water lakes. The purpose of this thesis is to find out if the conclusions drawn from the National water monitoring system are relevant to the drinking water industry based on the comparison of their catchment characteristics, thereby allowing the researchers and environmental analysts to use this data to draw conclusions concerning the Swedish drinking water. This was done by me creating a GIS shapefile of the catchments of lakes that are possible raw water sources. As a proxy for raw water sources we used lakes covered by water protection areas, as the location and ID's of the actual raw water sources aren't publicly available information. The attributes of these are then extracted and compared to the attributes of the National water monitoring system via non-parametric tests and comparison of the distributions. The aggregated forest characteristic was shown to be comparable to both NWMS categories, the peat depth, open wetland, longitude and water characteristic to the Omdrevslake category, and the latitude characteristic comparable to the Trendlake category.

Keywords: Catchment characteristics, National water monitoring system, raw water sources, Trendsjöar, Omdrevssjöar

Table of contents

List of Tables

List of Figures

Abbreviations

Abbreviation	Description
CC	Catchment characteristics
IQR	Interquartile Range
NMD	Nationell marktäckedata
NWMS	National water monitoring system
RWCC	Raw water catchment characteristic
RWS	Raw water sources
SLU	Sveriges lantbruksuniversitet
SMHI	Sveriges Meteorologiska och Hydrologiska Institut
SVAR	Svenska Vattenarkiv
TOC	Total organic carbon
VSO	Vattenskyddsområde

1. Introduction

The internal chemical composition of a body of water partly defines its ecological health through components such as nutrients, oxygen, pH and total organic carbon. The chemical makeup of a lake plays a critical role in determining its suitability for various uses, including recreation, agriculture and use as a water supply. These chemical attributes are all highly influenced by the surrounding land and its characteristics. Water that passes through a catchment ultimately ends up in a lake, transporting minerals and organic material that accumulate within the lake. Factors such as mineralogy, vegetation and agriculture all influence the water chemistry of bodies of water in the catchment this way(Kamenik et al., 2001).

A catchment of a lake is the geographical area surrounding a lake that encompasses all the areas from which water accumulates into said lake. All land where water travels downhill into streams, rivers and other bodies of water that ultimately end up in the lake is considered part of the catchment. Catchments are divided by topographic divides, such as hills or mountains, that direct surface water flow toward the lake.

The water chemistry of Swedish lakes have been monitored by The Swedish Agency for Marine and Water Management since 2006 in a program called the National Water Monitoring system, in groups of "Trendlakes" and "Omdrevslakes". The Trendlakes consist of 107 regularly sampled lakes that are specifically chosen to represent "reference conditions", i.e. conditions that are minimally impacted by human activity. Omdrevslakes consist of 5600 lakes of which 800 get surveyed each year on a rotating schedule. Omdrevslakes are "randomly" selected to represent the "average" Swedish lake, and show regional differences and status comparisons between Swedish lakes. (HaV, 2023, 2024)

The data collected from these lakes via the NWMS are widely used in research and assessments concerning the overarching state and changes in Swedish water, as the large time series and massive amounts of data provides a base for any research concerning boreal waters. topics such as brownification (Weyhenmeyer et al, 2014) and temporal trends (Von Brömssen et al 2023) to predict changes and potential threats to the water quality in Swedish lakes.

The Swedish drinking water industry is concerned with the chemical composition of the water used to produce drinking water. Using the conclusions drawn from research that utilizes the NWMS data concerning changes to the

Swedish drinking water would be very useful, and a great benefit to the drinking water industry.

The purpose of this thesis is to examine to what extent the data collected via the NWMS can be utilized to draw informed conclusions about the status and future of the Swedish raw water sources. This is done by comparing the catchment characteristics of Sweden's raw water source lakes with those of the lakes included in the NWMS.

As there is no readily available list of lakes used for drinking water in Sweden due to security reasons, a proxy of potential RWS lakes was constructed in this study. This is done by examining lakes that are protected by VSOs, which are water protection areas created primarily to protect drinking water sources from pollution and undesired human activity.

The catchments of the potential RWS will provide catchment characteristics of the RWS, attributes that show the surrounding land use. These attributes will be compared to the catchment characteristics of the Omdrevs- and Trendlakes, to find similarities and differences between the catchment characteristics, such as land use.

If the land use within the catchments of the RWS and NWMS lakes is similar, comparisons can be drawn between the 2 groups, as similar land use would imply similar chemistry within the lakes, at least when concerning the land use examined in this thesis. Similarity between the NWMS and RWS would allow an informed transfer of conclusions drawn from the water data provided by the NWMS surveys to be used to draw conclusions about the water chemistry within lakes used as raw water sources.

If the NWMS and RWS catchments turn out to not be statistically similar then the NWMS data cannot be used to draw conclusions about future changes in Swedish drinking water lakes. As land use and geographical attributes are local characteristics, extrapolating conclusions drawn from dissimilar characteristics would cause large uncertainties in the usefulness of the conclusions.

2. Method and materials

2.1 Data

To find out if the NWMS catchment characteristics are similar to the RWS catchment characteristics the attributes of the NWMS and RWS catchments need to be compared.

The catchment characteristics data of the Omdrevs- and Trendlakes are provided by the Department of Aquatic Sciences and Assessment at SLU. (https://miljodata.slu.se/MVM/)

The catchment data for the potential RWCC was extracted from shapefiles created in ArcGIS by me, using the SVARO catchments (SVAR2016) and Vattenförekomster (SVAR2022) shapefiles downloaded from SMHIs "Vattenwebb" database as well as VSOs (Naturvårdsverket, 2022) downloaded from Länsstyrelsernas Geodatakatalog database.

The Vattenförekomster consists of surface water lakes appointed by Vattenmyndigheterna as surface water bodies.
The SVARO catchments are accumulated from ARO catchments, small sub-catchments, created by SMHI for use in HYPE, their hydrological modelling system. These catchments contain information for topography, land use and soil types.
The VSOs are water protection areas created by the municipalities or counties that are within via hydrogeological assessment. VSOs are primarily created by the governing body to protect drinking water sources from environmental and restrict land use that might harm the quality of the water. Other bodies of water important to vulnerable ecosystems or used for recreation, irrigation and industry can also be protected via VSO, but it's not as common. The areas are compiled into a shapefile by Naturvårdsverket that show the affected area of these water protection areas.

The main attributes being compared between the RWS and NWMS catchments are:

1. NMD of the entire catchment (Land use)
2. Elevation of the outlet
3. Latitude & Longitude of the outlet
4. Density of ditches
5. Peat area

These shapefiles, a file format that stores geometric location and attributes, are used to represent the catchments of the **potential** raw water sources. To create a catchment shapefile that can be used to represent the catchments of potential water sources, lakes that are within a vicinity of 10 meters of a VSO are selected as potential raw water sources. These lakes are then matched with the closest catchment that fully contains them, creating a shapefile with all catchments connected to a potential raw water source. This results in us having a catchment for each individual lake that is overlapping a VSO.

The catchment shapefile is then exported from GIS and a python script is used to extract NMD, elevation, long & latitude, peat data and ditch area.

2.2 Processing of data

This data is then processed to be able to compare between sets, which includes normalizing ditch length and area as well as aggregating NMD according to documentation.

The data is represented as follows:

Attribute	Unit
Land Use	% of total catchment area covered by the land use characteristic
Area	Average area of catchments
Elevation	Meters above sea level
Latitude and Longitude	X and Y coordinates of SWEREF 99 TM
Ditches	Total length of ditches as m / ha of a catchments
Peat	Percentage of catchments area with peat depth of >=30 cm

Table 1: A table defining the attributes of the catchments being examined

The Land Use data (NMD) is a geographic database which describes the usage and coverage of land in Sweden, and is split into 7 different categories according to Naturvårdsverket documentation and classification of land use data. The categories are as follows:

1. Aggregated forest outside of wetland
2. Forested Wetland
3. Open Wetland
4. Arable Land
5. Other Open Land
6. Developed Land
7. Water

Elevation data along with the longitude and latitude of each catchment is used as is, based on the lowest point of the outlet in each catchment. The outlet elevation is used instead of the elevation in the center of the catchment, as the outlet should always be the lowest point in a catchment, whilst the central point of a catchment could be on an elevated area, which would misrepresent the actual elevation of the catchment / lake.

The length of ditches is taken from the SLU Ditch Map, normalized to the area of the catchment, and presented as m / ha.

Peat is shown as a percentage, representing the percentage of the catchment area that has a peat depth of at least 30cm. This is due to Riksskogstaxeringen defining peat as an area with 30 cm or more peat depth and SLU defining it as a class 1 peat depth. (Ågren & Lin 2022)

This along with elevation, peat depth, ditch area, longitude and latitude gives us 13 different parameters to compare. The catchment characteristic data from the raw water sources is compared to both the Omdrevs- and Trendlake catchment data, giving a total of 26 comparisons. These comparisons are made via Mann Whitney tests, a non-parametric test which results in a p-value that show whether there is a statistically significant difference between two groups. It is used since the samples are independent, the variance is high, and the samples aren't normalized. Mann Whitney tests that result in a p-value of 0,05 or higher, then the datasets can be assumed to not be significantly different.
The Mann Whitney tests also provide the r-value of the comparison, a standardized measurement of how strong the difference is between datasets.

The catchment characteristics are also processed into Box- and whisker plots, allowing easy viewing and comparison between the datasets, and provide further understanding of how similar or dissimilar they are outside of assumptions of statistical similarity. A table is also provided of the median of these plots, to provide an easier overview of median of these catchment characteristics.

3. Results

After constructing the catchments in GIS, a total of 419 were selected, with 13 of them being removed due to being incorrectly included in the method, as they were included despite being groundwater sources.

The data acquired from the Omdrevslakes consists of 5306 catchments, and the data acquired from the Trendlakes consists of 107 catchments.

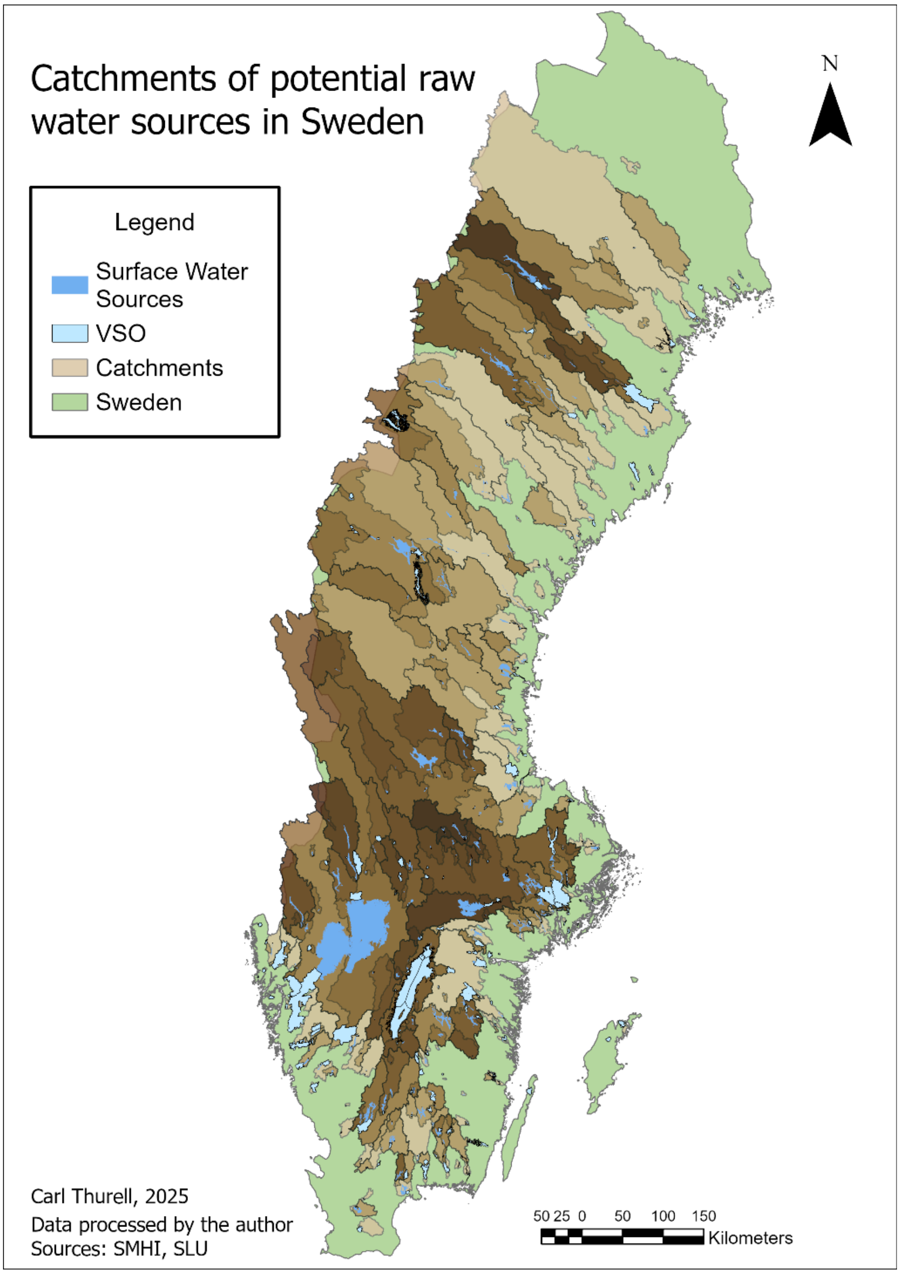


Figure 1: A map showing the catchments, lakes and VSOs of Swedish potential raw water sources

The map shows the placement of outlets and their catchments, the latter becoming darker as more catchments overlap. This reveals the density of catchment in certain areas

3.1 Mann-Whitney Tests

The 13 catchment characteristics of the Raw water sources were compared to the data of Omdrevslakes and Trendlakes via Mann- Whitney tests.

Characteristics	**Omdrevslake**	**Trendlakes**
	p-value	p-value
Area	0,00	0,00
Aggregated Forest	0,00	0,04
Forested Wetland	0,00	0,00
Open wetland	0,78	0,01
Arable Land	0,00	0,00
Open Ground	0,00	0,00
Developed Land	0,00	0,00
Water	0,00	0,00
Elevation	0,00	0,00
X-coordinates	0,07	0,99
Y-coordinates	0,00	0,39
Peat Area	0,91	0,00
Ditch Density	0,00	0,00

Table 2: The results of a Mann-Whitney test between the three lake categories: Raw water sources, Omdrevslakes and Trendlakes, and if the results show significant difference based on p-value.

The results show above are highlighted in green if they do not show significant difference, and red if they do show significant difference.

The Mann-Whitney test resulted in 21 out of 26 of the tests showing statistical difference. The 5 comparisons that showed no significant difference was:

1. Open Wetland between raw water sources and Omdrevslakes
2. The X-coordinates of raw water sources and Omdrevslakes
3. The X-coordinates of raw water sources and Trendlakes
4. The Y-coordinates of raw water sources and Trendlakes
5. The peat percentages of raw water sources and Omdrevslakes

Characteristics	Omdrevslakes	Trendlakes
	r-value	r-value
Area	0,3346	0,5165
Aggregated Forest	0,0464	0,0906
Forested Wetland	0,0512	0,1456
Open Wetland	0,0037	0,1223
Arable Land	0,2843	0,3741
Open Ground	0,0843	0,1907
Developed Land	0,1760	0,2981
Water	0,0474	0,1539
Elevation	0,2193	0,2656
X-Coordinates	0,0241	0,0005
Y-Coordinates	0,0653	0,0371
Peat Area	0,0015	0,1490
Ditch Density	0,3450	0,1683

Table 3: The results of a Mann-Whitney test between the three lake categories: Raw water sources, Omdrevslakes and Trendlakes, showing the R-value and the implied effect sizes (Cohen, 1988). Green cells show a small difference; yellow cells show a medium difference and red cells show a large difference

R-value is the standardized measure of how strong the difference is between two groups of data. The implied effect sizes are described by Cohen (1998) as rules of thumb of how a r-value represents the differences between 2 groups of data. An r-value of 0,30 or lower is considered small difference, a value between 0,30 and 0,50 is considered a medium difference, and a r-value of 0,5 or higher is considered a large difference.

The Mann-Whitney test resulted in r-values that show small effect size in 11 comparisons, medium effect size in 14 comparisons and large effect size in one comparison. The results are color coded, with small effect being green, medium being yellow and large being red. A lower r-value implies a small effect size, which in turn implies smaller practical differences between datasets.

Median	RWS	Omdrevslakes	Trendlakes
Area	194,1	2,7	5,7
Aggregated Forest	64,7	68,0	67,8
Forested Wetland	4,1	3,6	3,2
Open Wetland	3,4	3,6	2,8
Arable Land	2,8	0,0	0,0

Open Ground	4,6	2,9	2,1
Developed Land	2,6	1,5	1,7
Water	9,5	8,2	12,4
Outlet Elevation	81,3	189,5	164,2
Ditch Density	24,6	1,7	16,7
Peat Area	17,2	16,3	12,1
Longitude	537 737,0	526 700,9	538 945,9
Latitude	6 566 394,9	6 639 370,0	6 607 120,0

Table 4: A table showing the median of each compared catchment characteristic.

The median is calculated for each catchment characteristic instead of the mean to more easily visualize the size and percentage of the characteristics, as p-value and r-value only show distribution. When non-parametric tests such as Mann Whitney tests are used, the mean can be sensitive to extreme outliers, especially in larger datasets. The median is robust and won't be skewed by outliers, making it a better fit than mean

Table 4 and *Table* 5 shows that despite what the p-values shows, many median values are similar, especially between the RWS and Omdrevslake catchments.

	RWS / Omdrevslakes	RWS / Trendlakes
Area	1,39	2,96
Aggregated Forest	95,14	95,37
Forested Wetland	88,18	79,12
Open Wetland	93,06	82,45
Arable Land	0,00	0,45
Open Ground	62,32	45,82
Developed Land	58,12	63,93
Water	85,68	76,80
Outlet Elevation	42,90	49,50
Ditch Density	6,87	67,92
Peat Area	94,75	70,62
Longitude	97,95	99,78
Latitude	98,90	99,38

Table 5: The similarity of the median of catchment characteristics, shown in %

3.2 Box- and Whisker plots

The raw NMD data as well as the elevation, longitude and latitude, peat area and ditch area were also consolidated into box- and whisker plots. Shown here are the plots for area, aggregated forest, elevation, peat depth, open wetland, developed land, arable land and ditch density. The rest of the plots of the catchment characteristics can be found in the appendix.

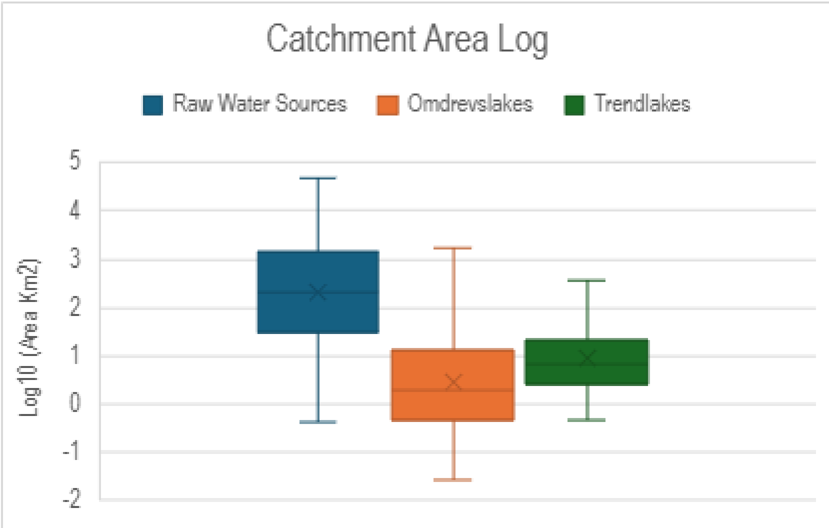


Figure 2: A Box- and whisker plot showing the distribution of the catchment Area of Raw Water Sources, Omdrevslakes and Trendlakes in log10

The area distribution of RWS is shown to be much higher on average. The RWS median of 194,079 km^2 is larger than the median of the Omdrevslakes (2,69 km$^{2)}$ and the Trendlakes (5,74km^2). Due to area data tending not to be normally distributed, the area is plotted in log^{10}. The RWS median and mean lie outside the box of the NWMS catchments at a higher value, implying that the average value of the RWS catchments is higher.

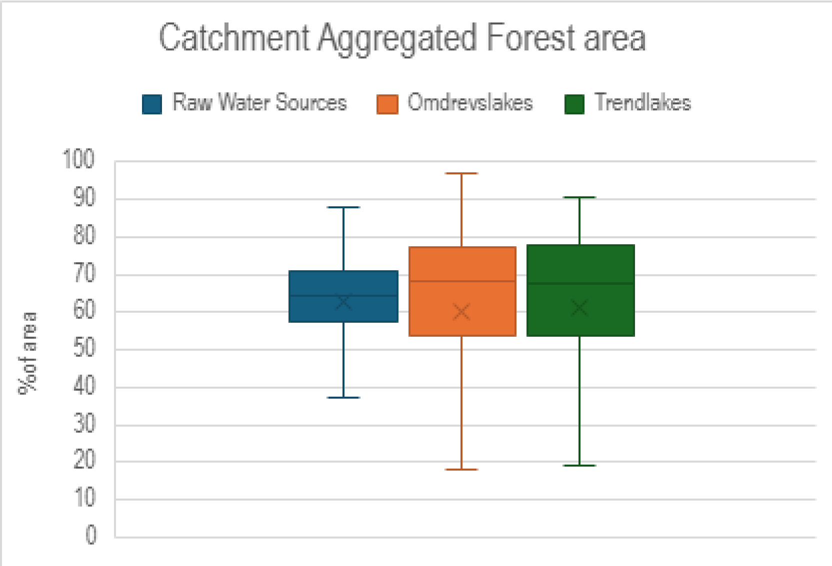


Figure 3: A Box- and whisker plot showing the distribution of the percentage of Aggregated Forested area in catchments of the Area of Raw Water Sources, Omdrevslakes and Trendlakes

The median between RWS only differs about 5% from both the Omdrevs- and Trendlakes and the r-value is 0,05 and 0,09 respectively, both small differences in effect size. This is also reflected in *figure 3*, where the RWS results look similar to the NWMS ones, only more concentrated. The medians and means intersect fully between all boxes, and the whiskers of the RWS are fully contained within those of the NWMS.

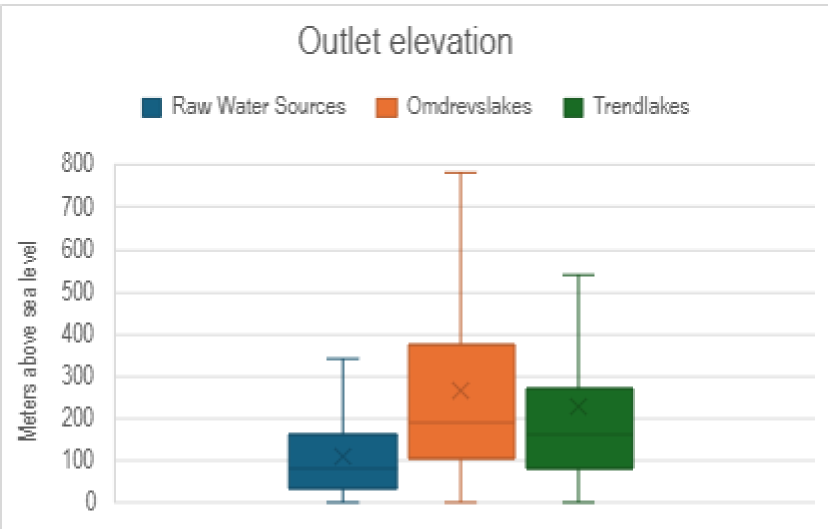


Figure 4: A Box- and whisker plot showing the distribution of the elevation of outlets connected to lakes from the Raw Water Sources, Omdrevslakes and Trendlakes

The RWS outlet elevation is shown to be lower than that of the Omdrevs- and Trendlakes. The range of the RWS catchments is also much lower, whilst the Trendlakes have a more even distribution and the Omdrevslakes have a positively skewed distribution. The max value of Omdrevslake is ~128% larger than the RWS and the max value of Trendlakes is ~58% larger, even though the RWS is fully contained within the Omdrevs- and Trendlake range, the distribution is not similar.

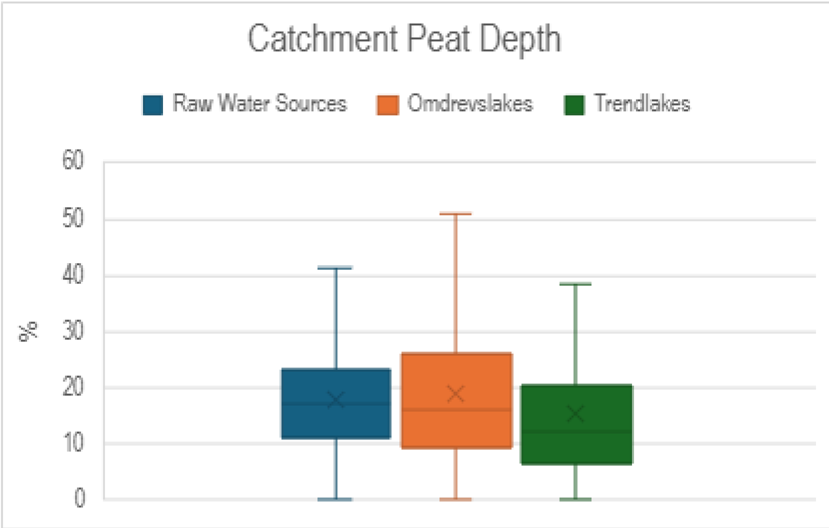


Figure 5: A Box- and whisker plot showing the distribution of the percentage of area with a peat depth of 30 cm or deeper in catchments of the Area of Raw Water Sources, Omdrevslakes and Trendlakes

The peat depth characteristic of the RWS and Trendlakes are similar in range whilst the Omdrevslakes have a larger range. The Trendlakes have a lower median and mean, whilst the median of the RWS and Omdrevslakes is similar, with the mean of the Omdrevslake being higher. The medians intersect all boxes, and the range of the RWS and Trendlakes are fully contained within the range of the Omdrevslakes.

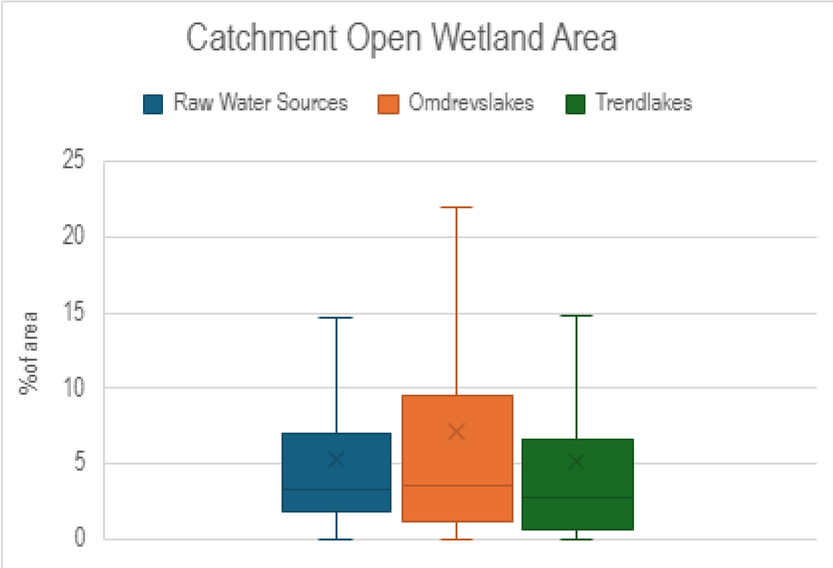


Figure 6: A Box- and whisker plot showing the distribution of the percentage of Open Wetland area in catchments of the Area of Raw Water Sources, Omdrevslakes and Trendlakes

The Open wetland results are similar, apart from Omdrevslakes having a larger spread and positive skew. The medians of the RWS and Omdrevslakes are similar, and the RWS data is fully contained within the spread of Omdrevslake data. The mean of the RWS and Trendlakes are similar, whilst the Omdrevslakes have a higher mean.

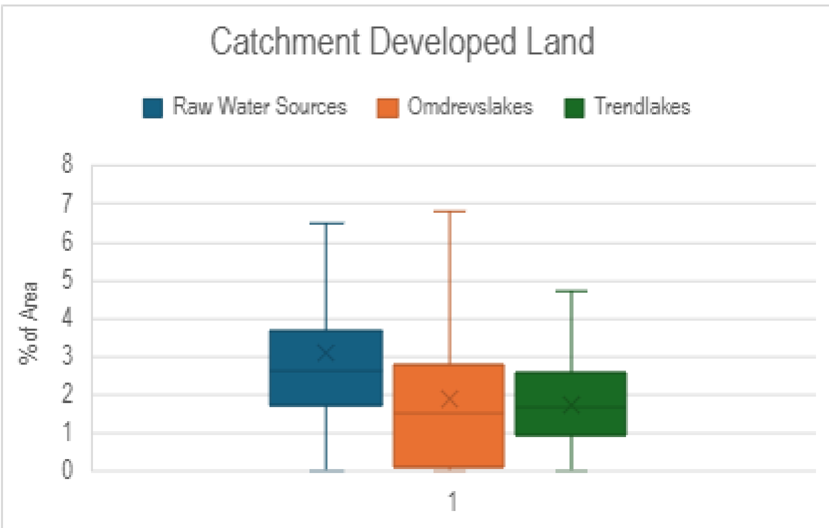


Figure 7: A Box- and whisker plot showing the distribution of the percentage of Developed land area in catchments of the Area of Raw Water Sources, Omdrevslakes and Trendlakes

The RWS developed land show a higher mean and median than the NWMS catchments. Whilst the range is similar between the RWS and Omdrevslakes, the

Trendlakes have a smaller range. The RWS results are completely within the results of the Omdrevslakes, but the Omdrevslakes are negatively skewed, making the distribution dissimilar.

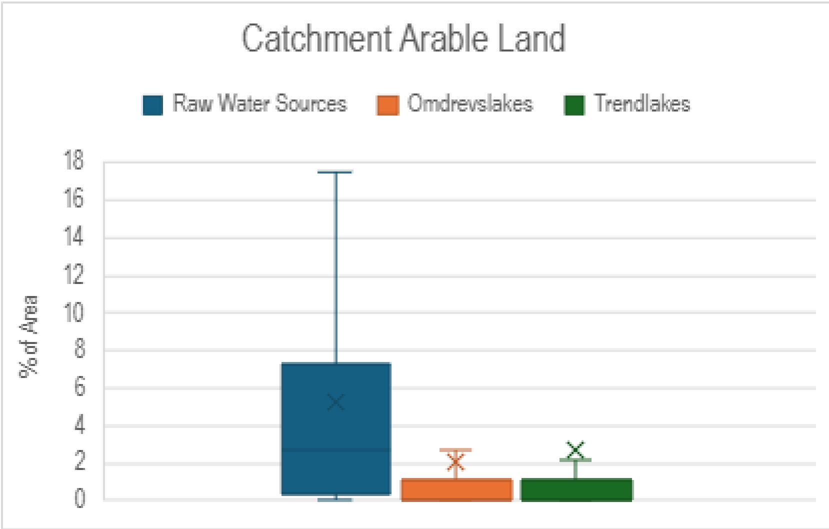


Figure 8: A Box- and whisker plot showing the distribution of the percentage of Arable land area in catchments of the Area of Raw Water Sources, Omdrevslakes and Trendlakes

Mean, median and range of the RWS is larger than the NWMS.

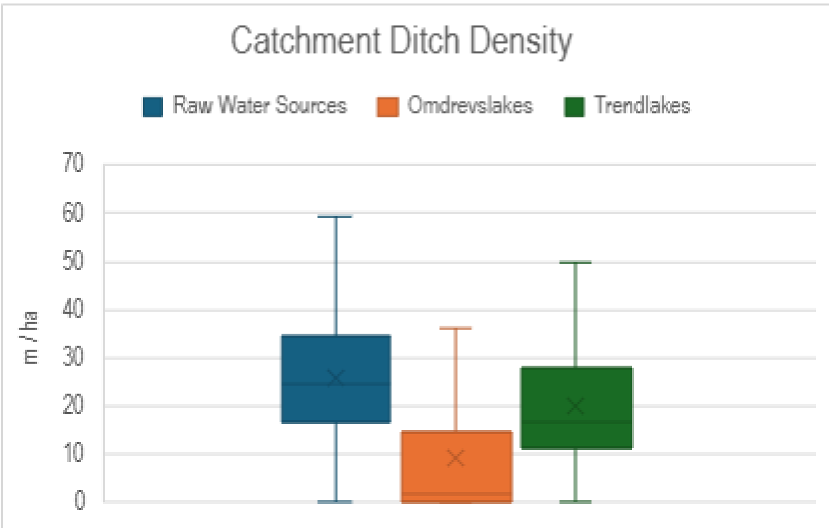


Figure 9: A Box- and whisker plot showing the distribution of m / ha of ditch area in catchments of Raw Water Sources, Omdrevslakes and Trendlakes

The ditch density results shows that the RWS and Trendlakes catchments both have somewhat similar characteristics, while the Omdrevslake catchments has a much lower median and average ditch density. Although some overlap between the RWS and Trendlakes is shown, the median and mean are different, making similarities unlikely.

4. Discussion

Before discussing how the results relate to the research question, characteristic comparison that showed unexpected or interesting results are discussed first, as it's easier to relate back to this discussion to draw conclusions about the research question.

4.1 The large area discrepancy

The biggest difference found between the RWS and NWMS is the area of the catchments, as can be seen in *fig 2, table 2 and table 3*. The median, which represents a typical value in the catchment characteristic, p-value, which shows if two compared groups are significantly different and r-value, which is the standardized measure of difference between 2 groups, all imply the largest difference of all the catchment characteristics compared. The most likely reason for this is the method used to select the NWMS lakes in comparison to the RWS lakes. The raw water sources are selected for one purpose alone, being a source of drinking water. This implies certain qualities, such as the necessary water balance to support constant water extraction, water quality and ease of access.

In smaller lakes these qualities are rarer, as a smaller buffer of water reduces the capacity to dilute the inflow of contaminants but also makes it vulnerable to periods of low water inflow (Moses et al, 2011). Smaller lakes are also limited in how much water can be extracted from the lakes due to smaller amounts of water inflow. All these factors result in smaller lakes generally not being chosen as surface water sources, at least not on an industrial scale.

Another explanation for this phenomenon is the SMHI sub-catchment system used to find potential raw water sources. Open data presented by SMHI, like the Vattenwebb data, is modelled using SMHI's hydrological model HYPE. This model uses a sub-catchment system, which causes smaller catchments to be grouped with adjacent ones to create larger catchments, which causes minor catchments around smaller lakes and water sources to not be represented in the HYPE model.
Despite the drawbacks, using the SMHI sub-catchments saves a lot of time and computing power compared to constructing catchments using digital elevation models.

This causes most smaller lakes to be omitted from the selection of raw water source lakes, while still being selected through the random selection method of Omdrevslakes and the specific selection method of Trendlakes, which look for other qualities such as lack of minimal local impact and representativeness rather than the access to drinking water (HaV, 2023).

4.2 Forested areas, elevation and the tree line

The aggregated forest data was the highest average percentage of the NMD data, according to *fig 3*, of the RWS and NWMS catchments, which is reasonable to assume since most of Sweden is covered in forest. (Wikberg et al, 2022)

With 28,1 million of Sweden's 40,7 million hectare being covered in forest (Roughly 69%) it's reasonable to assume that with a large enough sample size the forest coverage of Swedish lake catchments would be roughly the same. Despite this, the distribution of forested areas within Omdrev- and Trendlake catchments according to *fig 3* implies a wide range, with some areas having as little as 20% forest coverage, whilst the lowest coverage of RWS catchments is only slightly below 40%, resulting in a higher minimum forest cover for RWS catchments than NWMS lakes.

The most reasonable explanation for this is the forest line. As previously explained when discussing the area discrepancy, the criteria for selecting lakes for the RWS requires the practical application of extracting drinkable water from them. An area of Sweden where very few to no water source lakes and low forest coverage is the Swedish mountains. A large part of these areas is above the *Tree line,* a dividing line where trees no longer can grow. These areas start ranging from between 600 to 1140 meters above water, depending on temperature. *(Lindberg, 2019)*

Looking at *fig 4*, we can see that the max value of the RWS Outlet elevation never reaches these heights. Except for outliers, the RWS outlets only reach a height of about 350 meters above water, whilst the Omdrevslakes and Trendlakes reach around 800 and 550 meters above water respectively. It is also important to remember that the values shown in *fig 4* are the lowest flow point of each outlet, meaning that the surrounding catchment is very likely at a higher elevation than the given data.

Due to difficulties of extracting surface water in mountainous areas, the smaller population and the abundance of groundwater sources, many counties in these areas, such as Norrbotten and Västerbotten, are mainly focused on groundwater resources *(Länsstyrelsen Norrbotten, 2013;Länsstyrelsen Västerbotten, 2020)*. This would explain why no or very little RWS is taken from these areas, resulting in the range of Aggregated Forest being much more concentrated in RWS data, as it more closely follows the average Swedish forestation, due to the catchments of the RWS not being found in these mountainous areas with little to no forest coverage.

On the other hand, the max values of Omdrevs- and Trendlakes forestation is also higher than the RWS. This is most likely explained by the smaller catchments with smaller lakes, which allow for almost 100% forest coverage.

Although no significant similarity is shown in *table 2*, the r-value shown in *table 3* is very low. This means that in practice, the difference is very small, and when comparing larger datasets the p-value isn't always absolute as a low r-value in combination with a low p-value means that the results are statistically different, but practically similar. (Cohen, 1988).

This is further reinforced by *figure 3* which shows that the interquartile (IQR) range of the RWS data is highly concentrated and fully contained within the IQR of the NWMS data. This in addition to a similar meaning between the datasets would likely mean that catchments with aggregated forest inside the IQR of these datasets are comparable.

4.3 Wetland and Peat

Two of the attributes that have a high p-value and low r-value according to *table 2* and *table 3* are Open Wetland area (*fig 6*) and Peat Depth (*fig 5*) of the RWS and Omdrevslakes.

As peats accumulate in wetlands such as bogs, fens and marshes it's no surprise that a similar amount of wetlands would provide a similar amount of peat depth. Although similar p-values are shown between Open Wetland and Peat depth, the wetland coverage of the RWS (3,35% median) and Omdrevslakes (3,609% median) isn't anywhere close to the Peat depth coverage of the RWS (17,2% median) and Omdrevslakes (16,29% median). The two most likely reasons for this is that Forested wetland (*fig 12*) isn't included in the open wetland amount, leaving some wetland out of the observation and the change in land use over time.

Forested wetlands compose a median of 4,05% in the RWS catchments and 3,57% in Omdrevslakes, and whilst the p-value of forested wetlands is indicative of significant differences, the r-value of the comparison is very low, so it's reasonable that a practical similarity between them exists.

The other reason is the change in land use in Sweden. Since the middle ages, lowering of lakes and ditching of wetlands have been done to create cultivable land and promote forest growth in later days, whilst protection of wetlands is a relatively modern idea, starting in 1970s (Gunnar, 2019). The result of this is a large amount of wetland disappearing, being replaced by farmland or forest, but the layer of peat beneath it staying intact. This would explain the disparity in peat producing land in the catchments and actual peat amount, as much of the peat would've been created long before the forest and agricultural land use areas.

4.4 Proximity to settlements

The amount of developed land, which includes buildings, roads / railroads and exploited areas that aren't buildings, roads or railroads, is significantly higher in the RWS data than the NWMS data as shown in *table 2*.

A very likely reason for this is the historical placement of settlements. Historically, settlements have been placed near the coast or other large bodies of water, for a variety of reasons, including access to freshwater (*Groß et al, 2018).* This might facilitate a connection between the size of lakes, their use as freshwater sources and proximity to settlements.

But the opposite is likely also true, as freshwater sources would have been chosen simply due to the proximity of settlements. Many of our raw surface water sources used today aren't used due to superior water quality, but rather due to proximity and already established water sanitation industry (*Juuti et al, 2009).*

The Omdrevslakes are randomly selected, creating the possibility that many of the catchments include little to no developed land which can be seen in *fig 7* (HaV, 2023). In a similar vein, Trendlakes are selected with the absence of potential local disturbances in mind, creating catchments with a limited amount of developed land (HaV, 2024). Following this logic, it's easy to see why the catchments of lakes selected for their practicality and use for human consumption would include larger amounts of developed lands than catchments which do not.

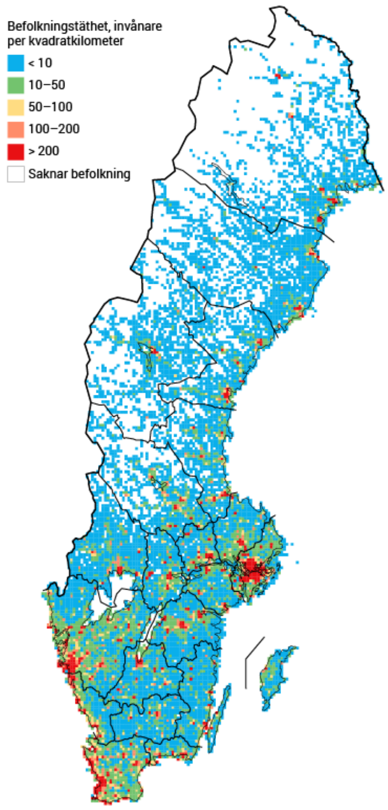


Figure 10: A map showing the population density of Sweden (SCB)

Another source of this statistical disparity is the latitude of the outlets of these datasets. As seen in *fig 15*, the average NWMS outlet is located north of the average RWS outlet. According to SCB the population density of Sweden is lower in the northern parts than the southern parts as seen in *fig 10*. This would suggest a correlation between average latitude of a catchment and the average developed land area of a catchment, with the higher latitude implying lower developed land area, which is consistent with the findings in amount of developed land area.

4.5 Agriculture and Ditches

According to *fig* 8, Arable land of the RWS has a much higher average than both Omdrev- and Trendlakes. This is most **likely** due to similar reasons as why the distribution of developed land in the catchments are different, that the higher latitude of the outlet, the lower the area covered by developed land.

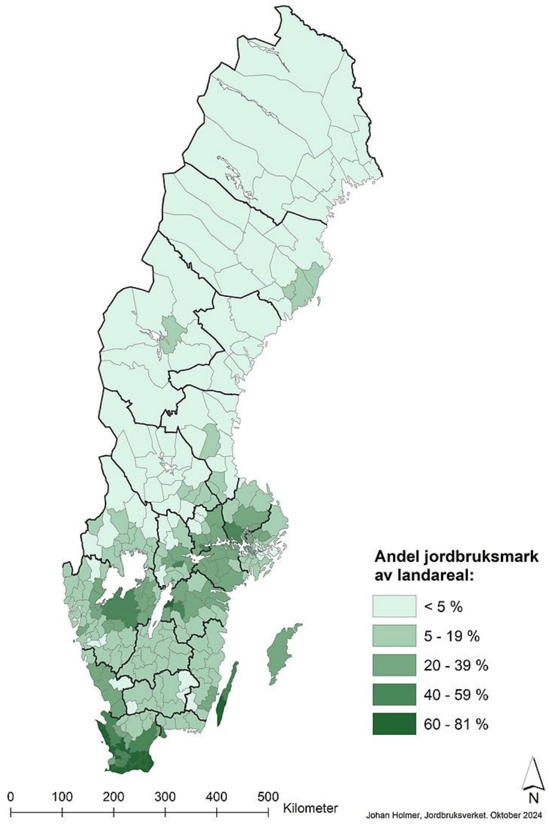


Figure 11: A map showing the proportion of arable land of Sweden (Jordbruksverket)

As can be seen in *fig 11*, the agricultural density follows population density, being higher in the south and central Sweden, whilst lower in the north. As previously discussed, a larger amount of RWS catchments can be found in the central and southern parts of Sweden as seen in *figure 1*, as opposed to the NWMS which are more equally spread.

Along with the possible connection between latitude and average arable land area of a catchment, the highest concentration of raw water source lakes is in Mälardalen. The average arable land area in Mälardalen is much higher than the average in Sweden, further raising the average arable land area of RWS catchments, even outside the general correlation with latitude.

Directly connected to this is the ditch area of the catchments. RWS catchments also have higher median in this attribute, although not by as large of a margin as with Arable land, and not as consistently either, as Trendlakes have a higher median distribution than Omdrevslakes, seen in *fig 9*.

The most common reason for creating ditches in Sweden is to drain wetland to create arable land (Gunnar, 2019). A possible explanation for this is that a large amount of arable land in an area would've most likely been preceded by a large amount of wetland, requiring ditching to transform the area. This explains why the RWS catchments have a higher ditch density than the NWMS catchments, but not why Trendlakes don't follow the correlation between arable land and ditch area.

This disparity could be caused by the selection process of Trendlakes. Trendlakes are used to monitor the chemical attributes of lakes without the influence of local factors. Reasonably, these lakes would not be placed within catchments with large amounts of arable land and agricultural industry, as these attributes would most likely influence the chemical composition of the water. Due to this, the ditching of Trendlake catchments might be due to other factors, such as the cultivation of forested areas, resulting in higher ditch density.

4.6 Are the catchments characteristics of the National Water monitoring system and the Raw water sources comparable?

Of the 26 catchment characteristic comparisons between the RWS and NWMS, only 5 had no significant difference. However, it is likely that the catchment characteristics are more comparable than indicated by the p-value.

As discussed earlier, interpretation of the box-and whisker plots along with a affirming r-value can give a better understanding of the similarity between the datasets than the p-value, as the p-value is only the measure of the null hypothesis. This is especially true in comparison with Omdrevslakes, as in larger datasets, a very small differences can result in low p-values, despite practical similarity between datasets.

Catchment characteristics that have intersecting medians and a large overlap in IQR and whiskers in the box- and whisker plots are the most interesting to look at, as the CCs without these qualities most likely aren't comparable between the RWS and NWMS.

An important note is that most similarities were found between the RWS catchments and the Omdrevslake catchments, while fewer similarities were found between the RWS and Trendlakes. The reason for this might be, as mentioned in the section just before, the selection process of the Trendlakes, which differ greatly from the criteria of RWS lakes, whilst Omdrevslakes are selected somewhat randomly. Although the sample size of Omdrevslakes would be larger and more likely to include the actual or similar lakes of the RWS lakes in its monitoring program, the Trendlake monitoring program actually has a larger percentual overlap between lakes in the Trendlakes than the Omdrevslakes, as seen in *table 5*.

	Amount of RWS lakes contained within the NWMS monitoring system	% of lake overlap
Omdrevslakes	183	3,44
Trendlakes	5	4,71

Table 6: The amount of RWS lakes contained within the system of Omdrevs- and Trendlakes

After comparing the p-values, r-values and box plots, only a few characteristics showed evidence of similarities large enough to discuss.

The catchment characteristics that exhibit the greatest similarities and may therefore be considered comparable include:

1. Aggregated Forest
2. Peat Depth
3. Open Wetland
4. Longitude & Latitude
5. Water

The aggregated forest characteristics show many similarities between the RWS and both NWMS categories. Whilst the p-values show a significant difference, the r-values and Box- and whisker plots show similar results.

Due to this, it's reasonable to draw conclusions about the forest characteristics of the RWS catchments using data collected from the NWMS catchments, as long as it doesn't concern extreme outliers in aggregated forest amounts.

Peat depth comparison between the RWS and Omdrevslakes catchments also show very similar results. The p-value implies no significant difference, the r-value is small, and the medians are similar. The box plot distribution is also very similar, although the RWS is has a lower max value, making it more concentrated, but only slightly so. This catchment characteristic exhibits a high degree of similarity, making it feasible to use peat data from the Omdrevslakes catchment for drawing informed conclusions about the RWS catchments.

In a similar vein, the Open wetland characteristic of the RWS and Omdrevslake catchments are also similar. The p-value again shows no significant difference, and the r-value is even lower. The box plot is like the Peat depth one, with the distribution being slightly more unevenly skewed for the Omdrevslakes. Whilst the mean isn't similar, the median is similar, making comparisons between these characteristics very feasible.

The longitude characteristic is similar between all the lake categories, as seen in *figure 16*, which signifies that the placement of lakes running east to west is similar between all categories of lakes. Latitude is comparable between the RWS and Trendlakes. This implies that the lakes are reasonably alike in the placement from South to North, which in Sweden can have major implications on the environment, ecology, flora and fauna and more (Callaghan et al., 2013). The p-value, r-value and box plot all show that conclusions about latitudinal effects of the RWS catchments can be drawn from the Trendlake data.

Lastly, the Water characteristics show some similarities as well. The box plot (seen in *figure 14*) are arguably the most similar of all the characteristics, showing only very slight differences between the RWS and Omdrevslake data. Similar to the Aggregated Forest comparisons, the p-value implies significant difference, the r-value implies smaller differences. The difference in median is slightly larger in this case, making it the least comparable of the 4 characteristics recounted here, but still a viable characteristic to compare between the RWS and Omdrevslake catchments.

The other catchment characteristics discussed in this thesis lack strong enough similarity to be useful in comparison.

4.7 Limitations

Potential RWS

A large source of error in this thesis is the fact that there is no complete list of all Swedish raw water sources. Part of creating this thesis was creating a catchment layer in GIS with an approximation of the Swedish raw water sources.

The matching of lakes within a proximity of VSOs to catchments resulted in this layer, and some manual filtering was done when very large VSOs were found to be matched to very small lakes, as these would often be groundwater sources that happened to be close to a lake that's protected by a VSO.

As such, this thesis can only give an estimate result of the characteristics of raw water source catchments.

Time restraints

To save time and computational power the catchments were created using SMHI sub-catchments.

The difference between using catchment delineations made with the exact outlet point of the lake using a DEM (Digital Elevation model) and the ARO sub-catchments is the minimum size of catchments. As mentioned earlier, sub-catchments that are small enough are joined to nearby catchments, causing smaller lakes to be omitted from the method.

Manual removal or inclusion of Catchments

In creating the GIS layout of the RWCs, there were two possible sources of error.

The first one was in the selection method. The method used would select every lake within the VSO buffer and fully contained within a SVARO catchment. This worked well, except for two lakes that consistently were not selected by the method, despite multiple retries. The first one, Svensbyfjärden, had a tiny part of its area overlapping into the ocean, which is reasonable since it's connecting with the ocean. But the SVARO layer used doesn't cover any ocean, meaning that this tiny area made the lake not fully contained within a SVARO catchment, causing the method to omit this lake.
The other lake was Lille Dalevatten, which for some strange reason overlapped with multiple SVARO catchments but wasn't fully covered by any of them.

Both these lakes were manually added. The most likely reason for this mismatch of SVAROs and lakes is the different edition of datasets. The SVARO layer used is from 2016, as there hasn't been a SVARO layer of Sweden created later, whilst the VSO and water layers are both from 2022. This might have caused a disparity between the water and SVARO layer, causing this hiccup.

The second problem was due to the polygonal composition of the lakes. Many of the lakes in the water layer used are composed of multiple polygons. These polygons have separate MS_IDs, meaning that even though they have the same name, they will be treated as different lakes. This caused the same catchment to be selected for the same lake multiple, but for different parts of the lake. Using the same catchment multiple times for each lake with multiple parts wouldn't be completely inaccurate, but the method gets complicated if there is a balancing act in deciding what lakes are large enough to include multiple water sources.

Most notably many of the larger lakes are composed of multiple parts, such as Vänern, Vättern, Mälaren and others. This would cause great inaccuracies in the results since these lakes have huge catchments that would balloon the average area of the RWSCC. Due to this, all lakes consisting of multiple polygons were joined to a single polygon.

Mann Whitney U-test vs t-test

A large part of this thesis is the statistical comparison of the RWS data and NWMS data. T-tests remain robust even in larger sample sizes, which would be useful considering the number of datapoints contained in the Omdrevslakes (ca 5300). In contrast, Mann Whitney tests are increasingly sensitive to small differences the larger the dataset is, resulting in low p-values despite highly similar datasets.

The issue with t-tests is that they assume normality and some amount of equal variance. This is less necessary with larger datasets, but skewed data and extreme outliers can heavily influence the result, which happened to be the case with many of the datasets used datasets. In these cases that lack normality, Mann Whitney tests are generally preferred. (Zimmerman, 1987)

This resulted in a unique problem. The datasets consist of RWS which are ca 400 datapoints which aren't normally distributed and have a high variance, Omdrevslakes, which are ca 5300 datapoints which aren't normally distributed and have a high variance and Trendlakes, which are ca 100 datapoints which aren't normally distributed and have a high variance. Whilst a t-test would probably provide better results when analyzing the distribution of Omdrevslakes, the smaller sample sizes of RWS and
Trendlake data would be better suited for Mann Whitney tests. In the end, the decision to use Mann Whitney tests were made. This was due to the availability of the r-value, which in combination of Box- and whisker plots allows for interpretation despite the result of the p-value, especially when comparing the RWS and Omdrevslake datasets.

4.8 Conclusions

This thesis evaluated whether conclusions drawn about the impact of catchment characteristics on water chemistry from the NWMS apply to Swedish raw water source lakes. Based on statistical testing and visual comparison of catchment characteristics of the RWS and NWMS, this thesis finds that the NWMS data can be used to draw informed conclusions about raw water source catchments in Sweden, provided that the data used is related to characteristics with that were shown to be comparable, which include aggregated forest area, peat depth, open wetland, longitude, latitude and water area. However, as these conclusions were drawn from results provided by proxies of the raw water sources, the results are only estimates and not absolute.

The catchments characteristics mentioned above show sufficient similarity to use in drawing broader conclusions, while other characteristics should be interpreted with caution due to noticeable dissimilarities.

The use of both statistical and visual analysis proved valuable in reaching an informed evaluation, as significant testing in combination with observational statistics in combination gave a more comprehensive overview of the similarities between the RWS and NWMS.

The Omdrevslakes catchment characteristics proved to be more representative of the RWS catchments than the Trendlakes, making them more suitable for further analysis in the future, and most likely more valuable when it comes to cross-use of data to draw informed, data-based conclusions about the state and future of the Swedish drinking water lakes.

References

Havs- och vattenmyndigheten. (2023). *Delprogram trendstationer sjöar – Programbeskrivning* (Rapport Nr. 2023:13).

Havs- och vattenmyndigheten (2024) *Delprogram Omdrevsstationer sjöar – Programbeskrivning*. (Rapport Nr 2024:4).

Ågren, A. M. & Lin, Y., 2022. *Dokumentation nya torvkartor – klassade torvkartor och kontinuerliga kartor över organiska lagrets tjocklek*. Version 1.0. Sveriges lantbruksuniversitet.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. 2:a uppl. Hillsdale, NJ: Lawrence Erlbaum Associates.

Moses, S.A., Janaki, L., Joseph, S., Justus, J. & Ramakrishnan Vimala, S., 2011. *Influence of lake morphology on water quality*. Environmental Monitoring and Assessment, 182(1–4), pp.443–454.

Walter, J., Fleck, R., Pace, M.L. & Wilkinson, G. (2020). Scaling relationships between lake surface area and catchment area. *Aquatic Sciences*, 82, 55.

Wikberg, P.-E., Roberge, C., Dahlgren, J., Fridman, J., Nilsson, P., Markström, M. & Borin, O., 2022. *Skogsdata 2022*. Sveriges lantbruksuniversitet.

Johansson, Julia (2019)**.** *Den alpina trädgränsen i ett föränderligt klimat – Observationer från svenska fjällkedjan*. Kandidatuppsats, Lunds universitet.

Länsstyrelsen Västerbotten**.** (2020). *Dricksvattenförsörjning – regional vattenförsörjningsplan för Västerbottens län*.

Sobek, S., Nisell, J. & Fölster, J., 2009. *Prediktion av sjödjup och sjövolym från topografiska data*. Rapport 2009:19. Uppsala: Institutionen för vatten och miljö, Sveriges lantbruksuniversitet (SLU).

Länsstyrelsen Norrbotten. (2013). *Regional vattenförsörjningsplan för Norrbottens län*.

Gunnar Jacks (2019) *Drainage in Sweden -the past and new developments*, Acta Agriculturae Scandinavica, Section B — Soil & Plant Science, 69:5, 405-410

Groß, D. *et al.* (2018) 'People, lakes and seashores: Studies from the Baltic Sea basin and adjacent areas in the early and Mid-Holocene,' *Quaternary Science Reviews*, 185, pp. 27–40.

Juuti, P. *et al.* (2009) *'Shared history of water supply and sanitation in Finland and Sweden, 1860-2000*,' *Vatten*, 65(3), pp. 165–175.

Statistiska centralbyrån (SCB). (n.d.). Befolkningstäthet i Sverige. *SCB – Sverige i siffror*

Jordbruksverket. (2024). *Jordbruksmarkens användning 2024 – slutlig statistik*.

Zimmerman, D.W. (1987) 'Comparative power of StudentTTest and Mann-WhitneyUTEST for unequal sample sizes and variances,' *The Journal of Experimental Education*, 55(3), pp. 171–174.

SMHI. (n.d.). *Ackumulerade uppströmsområden (SVAR2016)* https://www.smhi.se/data/sok-oppna-data-i-utforskaren/ackumulerade-uppstromsomraden-svar2016

Weyhenmeyer, G.A., Prairie, Y.T. and Tranvik, L.J. (2014) 'Browning of Boreal Freshwaters Coupled to Carbon-Iron Interactions along the Aquatic Continuum,' *PLoS ONE*, 9(2), p. e88104. https://doi.org/10.1371/journal.pone.0088104.

Von Brömssen, C., Fölster, J. and Eklöf, K. (2023) 'Temporal trend evaluation in monitoring programs with high spatial resolution and low temporal resolution using geographically weighted regression models,' *Environmental Monitoring and Assessment*, 195(5). https://doi.org/10.1007/s10661-023-11172-2.

Naturvårdsverket. (2022). *Skyddade områden, vattenskyddsområden*. Länsstyrelsernas Geodatakatalog. Tillgänglig vid: https://ext-geodatakatalog-forv.lansstyrelsen.se/PlaneringsKatalogen/GetMetaDataById?id=ae8d79d2-a799-4e1b-b500-05747a428816_C

Naturvårdsverket (2006). *Naturvårdsverkets föreskrifter om klassificering och miljökvalitetsnormer avseende ytvatten*. NFS 2006:1. Stockholm: Naturvårdsverket.

Crapart, C., de Wit, H. A., Austnes, K., & Hessen, D. O. (2023). *Spatial predictors and temporal forecast of total organic carbon levels in boreal lakes*. Science of The Total Environment, 869, 161676.

Callaghan, T.V., Björn, L.O., Chapin III, F.S., Chernov, Y., Christensen, T.R., Huntley, B., Ims, R.A., Johansson, M., Jolly, D., Jonasson, S., Matveyeva, N., Oechel, W., Panikov, N., Shaver, G., and Sitch, S., 2013. *Ecosystem change and stability over multiple decades in the Swedish subarctic: complex processes and multiple drivers*. Ambio, 42(S1), pp.10–23. https://doi.org/10.1007/s13280-012-0365-6

Sveriges lantbruksuniversitet, utan år. *Miljöövervakning – Data från miljöövervakningsprogram (MVM)*. Tillgänglig på: https://miljodata.slu.se/MVM/

Kamenik, C. *et al.* (2001) 'The influence of catchment characteristics on the water chemistry of mountain lakes,' *Arctic Antarctic and Alpine Research*, 33(4), pp. 404–409. https://doi.org/10.1080/15230430.2001.12003448.

Appendix

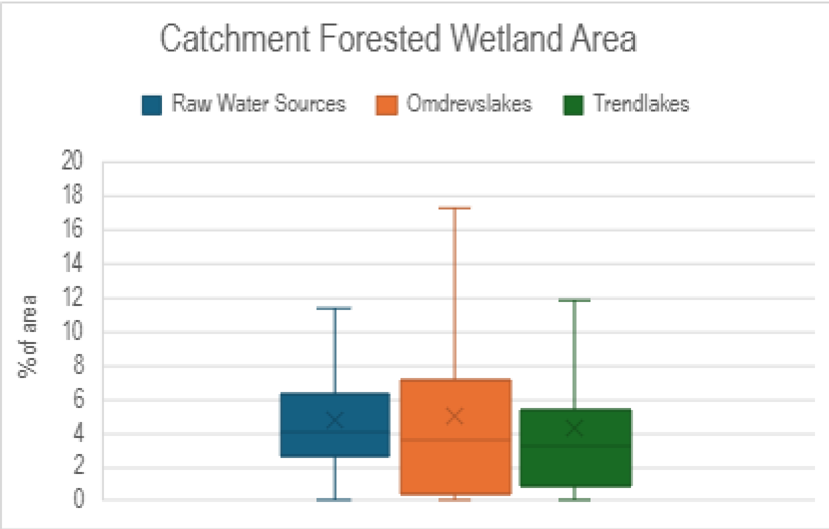


Figure 12: A Box- and whisker plot showing the distribution of the percentage of Forested Wetland area in catchments of the Area of Raw Water Sources, Omdrevslakes and Trendlakes

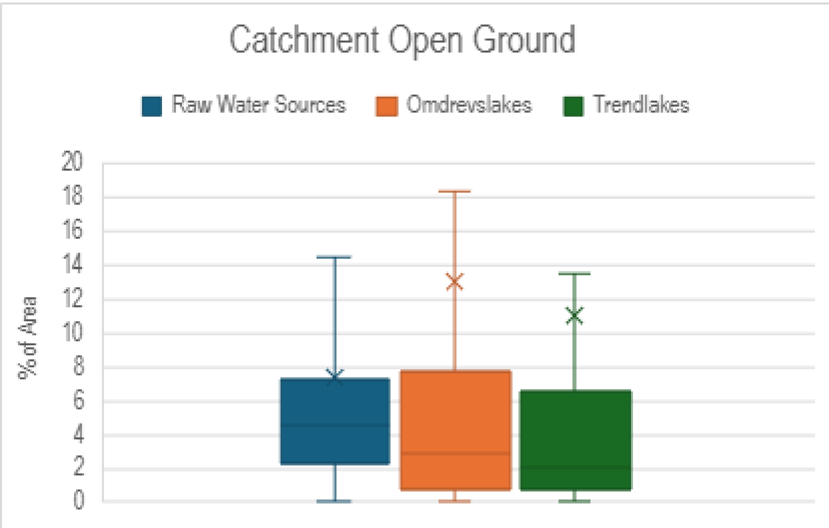


Figure 13:A Box- and whisker plot showing the distribution of the percentage of Open Ground area in catchments of the Area of Raw Water Sources, Omdrevslakes and Trendlakes

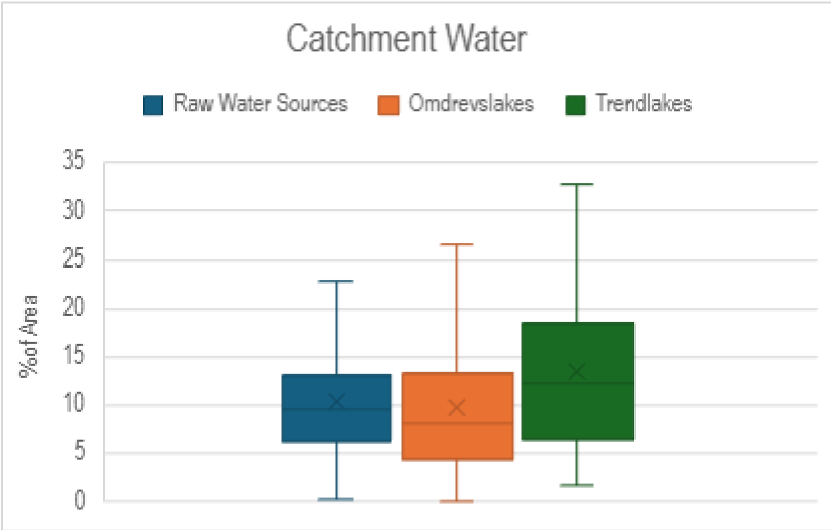


Figure 14: A Box- and whisker plot showing the distribution of the percentage of Water area in catchments of the Area of Raw Water Sources, Omdrevslakes and Trendlakes

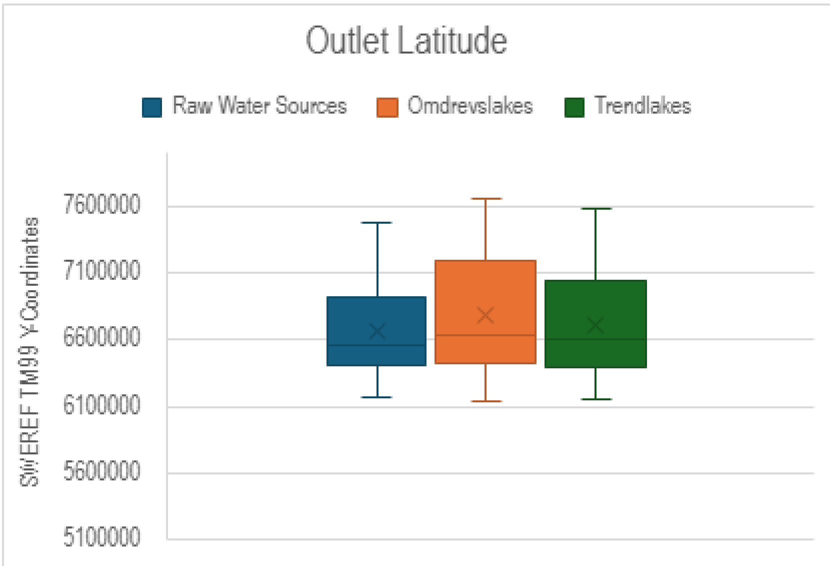


Figure 15: A Box- and whisker plot showing the distribution of latitude of outlets in catchments of the Area of Raw Water Sources, Omdrevslakes and Trendlakes

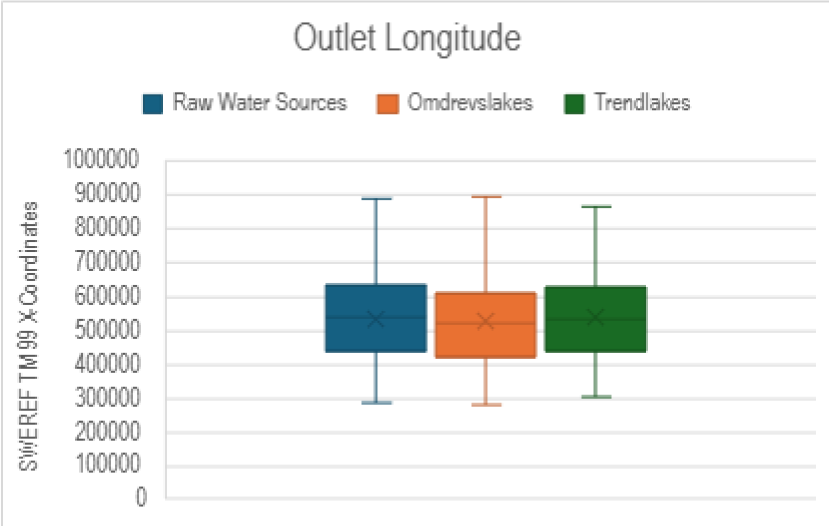


Figure 16: A Box- and whisker plot showing the distribution of longitude of outlets in catchments of the Area of Raw Water Sources, Omdrevslakes and Trendlakes

	p-value	p-value
Area	4,59E-04	2,4648E-37
Aggregated Forest	0,0004588480	0,04033534105
Forested Wetland	0,0001107750	0,00098875465
Open wetland	0,7789588249	0,00565907856
Arable Land	0,0000000000	0,00000000000
Open Ground	0,0000000002	0,00001589715
Developed Land	0,0000000000	0,00000000002
Water	0,0003433025	0,00049623824
Elevation	0,0000000000	0,00000000125
X-coordinates	0,0697045166	0,99060506701
Y-coordinates	0,0000008531	0,39476004248
Peat Area	0,9116300926	0,00064116663
Ditch Area	0,0000000000	0,00161396781

Table 7: A table containing the p-values of the Mann Whitney test results

Publishing and archiving

Approved students' theses at SLU can be published online. As a student you own the copyright to your work and in such cases, you need to approve the publication. In connection with your approval of publication, SLU will process your personal data (name) to make the work searchable on the internet. You can revoke your consent at any time by contacting the library.

Even if you choose not to publish the work or if you revoke your approval, the thesis will be archived digitally according to archive legislation.

You will find links to SLU's publication agreement and SLU's processing of personal data and your rights on this page:

- https://libanswers.slu.se/en/faq/228318

☒ YES, I, Carl Thurell, have read and agree to the agreement for publication and the personal data processing that takes place in connection with this