statvis Trust Through Data

Alberta Background Soil Quality System
Supporting Technical Document for Identification of Background Salinity Data Records and Ranges
Phase 1: Prototype for Pilot Area

Report prepared for
InnoTech Alberta

January 19, 2023

Table of Contents

Appendices

1. Introduction

Phase 1 of the Alberta Background Soil Quality System Project (ABSQS) is funded by InnoTech Alberta (InnoTech), the Alberta Upstream Petroleum Research Fund (AUPRF) managed by the Petroleum Technology Alliance of Canada (PTAC), and the Clean Resources Innovation Network (CRIN). The objective of the ABSQS is to develop a database of background metals and salinity parameters in the Province of Alberta for the purpose of decreasing the cost and time required to identify and remediate valid contaminants of potential concern (COPCs) on contaminated sites. The ABSQS is currently in Phase 1 which involves developing and testing the database for a pilot area. Acquisition of high-quality soils data is key to the ABSQS's overall success and to that end, owners of contaminated sites provided data from their sites under confidential data sharing agreements with InnoTech.

2. Background, Objectives, and Scope

2.1 Background

Salinity and metals parameters are some of the most common naturally elevated parameters in Alberta. Industry, government, and environmental consultants have identified a need for more effective identification of background salt and metals concentrations. There is currently no publicly available resource that maps or predicts background concentrations of these parameters for the Province of Alberta.

2.2 Objectives

Statvis Analytics Inc. (Statvis) was contracted by InnoTech to harmonize and clean the data, as well as develop a workflow to remove impacted soil data records leaving only background soil data records in the ABSQS database. The objective of the overall project is to work collaboratively with numerous actual and potential users of background soil data to develop the ABSQS. The ABSQS is intended to be used as a resource to assist industry and government in environmental management of contaminated sites. The objective of the current phase (Phase 1) of the project is to create a prototype version of the ABSQS for a pilot area. The pilot area was chosen based on the density of data provided in the various datasets. Starting with a pilot area vs. provincial-scale area is advantageous because:

- it allows the project team to test workflows,
- ensures stakeholder feedback can be incorporated before expanding to a provincial scale, and
- builds relationships with both data providers and users of the system.

In subsequent project phases, predictive mapping technologies will be applied to this dataset to create relevant spatial predictions of salinity and metals values across Alberta.

2.3 Scope of Work

For Phase 1 of the ABSQS, Statvis is responsible for compiling, harmonizing, and cleaning soil salinity and metals data into a geodatabase for subsequent analysis and use in predictive mapping platforms.

This document describes the workflow followed to complete these objectives for salinity data records within the initial pilot area defined in Figure 1. A subsequent report will be created for metals.

Salinity parameters of interest comprised calcium, chloride, magnesium, potassium, sodium, sulphate, electrical conductivity (EC), sodium adsorption ratio (SAR) and pH. These nine parameters were selected as they are most commonly reported in salinity analytical packages across analytical laboratories. These parameters are commonly analyzed because they have potential to affect the ability of soil to support plants and soil microbes as well as soil structure. They are also the most relevant parameters for complying with regulatory obligations and separating multiple salinity sources in the Province of Alberta.

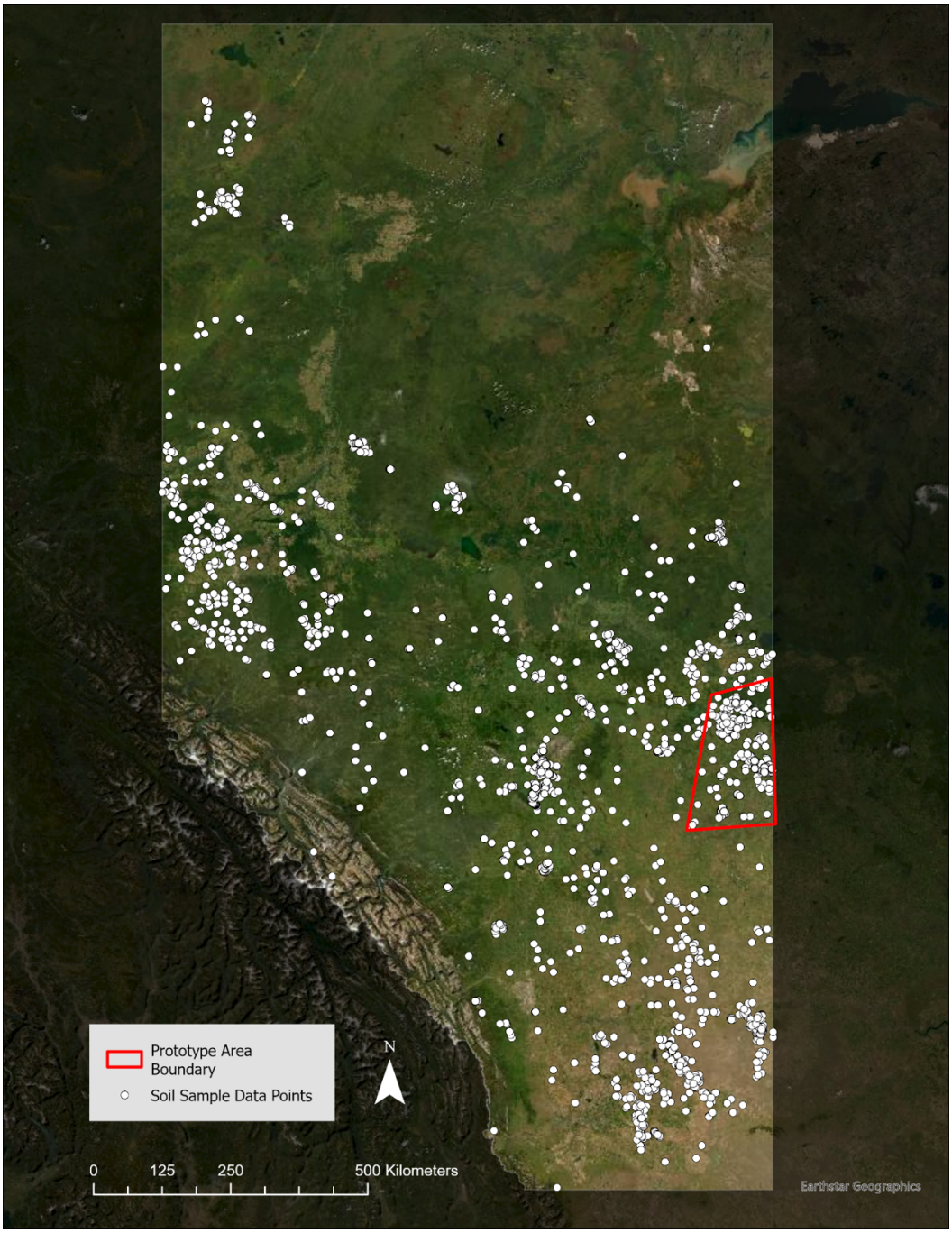


Figure 1. Full Salinity dataset and pilot area (labelled Prototype Area Boundary) data record sites.

3. Methodology

To ensure creation of a high-quality dataset that reliable conclusions could be drawn from, datasets were collected and then prepared, explored, and analysed according to the methods described below. Multiple workflows were tested. The chosen workflow proved the most effective for separating data records with anthropogenic influence from data records representative of background conditions.

3.1 Data Compilation

During the first eight months of the project (i.e., August 2021 to April 2022), Statvis engaged with data providers to request data and ensure data received was formatted correctly. The identity of data providers and details about the quantity and types of data provided are protected as confidential under data sharing agreements. In several instances there were formatting issues or missing metadata (e.g., UTM zones, units, or analytical methods). To resolve these issues, Statvis engaged with the data provider, and the dataset was either re-exported to correct the issue or information was provided so that Statvis could manually correct the issue. Over 2,700 individual data files from eight different data owners were received for the pilot area.

3.2 Data Harmonization

Harmonization of datasets was then carried out. Harmonization is the process of combining multiple smaller datasets into one master dataset. Columns were matched based on parameters and metadata values and combined to create a master dataset for the project comprising 23,821 records.

3.3 Data Cleaning

The master dataset was cleaned to remove incorrect, erroneous, or duplicated data. The dataset was first limited to only those data records where all nine salinity parameters of interest were present. Data records with missing parameters of interest were excluded from the initial analysis. It is important to note that these data records were not deleted, simply excluded from the initial analysis. These steps removed 143 data records, leaving 23,678 data records in the analysis-ready salinity dataset for the pilot area.

3.4 Data Exploration and Dimensionality Reduction

Statvis used hierarchical cluster analysis (HCA) alongside traditional statistical techniques (correlation plots and summary statistics for various subsets of the data) to identify clusters of data records representative of anthropogenic and non-anthropogenic (i.e., background) patterns. Clustering in general is a method of statistical analysis that clusters data records in such a way that they are more like other data records within the same cluster than they are to data records in other clusters. HCA is used to find discrete clusters with varying degrees of similarity (or dissimilarity) in a dataset. HCA builds a hierarchy of clusters and displays them on a dendrogram. A dendrogram is a tree-structured graph that shows the relationship between data records based on the length of the line connecting them. Shorter lines represent a closer relationship while longer lines indicate a larger difference between data records. See Figure 2 for an example HCA dendrogram. As distance from individual data records increases dendrogram lines become longer showing more dissimilarity between data records. In the example shown in Figure 2 three lines have been drawn bisecting the dendrogram, labeled y1, y2, and y3 respectively, that show three options for clustering granularity. Line y1 splits the dataset into 10 clusters, y2 splits the dataset into seven clusters and y3 splits the dataset into four clusters. As the number of clusters increases, the relationships between individual data records in a cluster become more granular and specific. For example, line y1 provides so much granularity that several the clusters have only one data record in them.

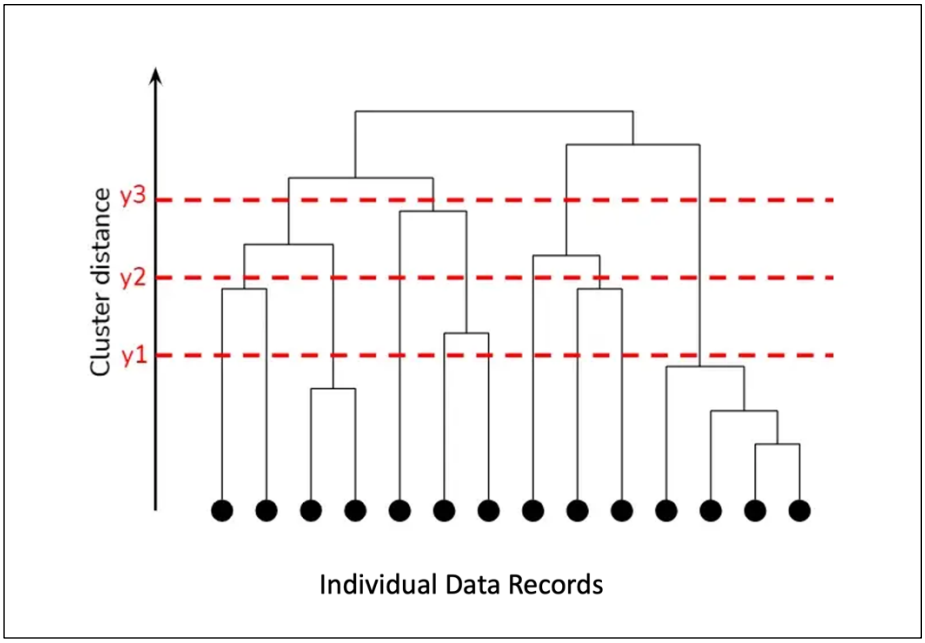


Figure 2. Example HCA dendrogram with bisecting lines to show various levels of granularity.

The goal of exploring the ABSQS salinity dataset using HCA was to provide enough granularity that data records showing anthropogenic impacts could be separated from clusters representative of background conditions. To achieve this, boundary conditions for an ideal background dataset had to be defined.

3.5 Establishing an Ideal Background Dataset

To ensure a high degree of confidence in removal of impacted data records from the master background dataset for the pilot area, a conservative approach was used to select the ideal background dataset. For the purposes of the project, an ideal background dataset was defined as:

- containing only data records having detectable concentrations of parameters being used to establish background patterns to ensure that relationships between all parameters could be examined;
- data records that have been identified as background with a high degree of confidence based on factors including, but not limited to, location (i.e., offsite), professional judgement, and concentration limits (i.e., guidelines or applied heuristics); and
- containing enough data records to allow reliable patterns to be identified.

The subsoil salinity tool manual states that background chloride concentrations "...vary between different regions and soil types but are generally below 100 mg/kg…".[1] This heuristic is often applied by provincial regulatory agencies and is considered very conservative. Conservatism was applied to ensure a high degree of confidence in the background patterns identified and used to remove impacted data records from the full pilot area dataset.

Approximately 6,000 data records of the 23,678 data records in the analysis-ready dataset from the pilot area had over 100 mg/kg chloride. This dataset was further reduced by removing data records that had either:

- non-detectable values for one or more of calcium, chloride, magnesium, potassium, sodium, sulphate, and EC; or

[1] Equilibrium Environmental 2020. *Subsoil Salinity Tool Version 3.0 User Manual.*

- an ionic imbalance of greater than 25%.

Data records with non-detectable values of one or more of the six salt ions (calcium, chloride, magnesium, potassium, sodium, and sulphate) or EC were removed to ensure that definitive patterns were able to be developed based on relationships between these important salinity parameters. SAR and pH values were not mandatory for inclusion in the ideal background dataset as SAR is calculated from sodium, calcium, and magnesium concentrations and pH did not provide diagnostic relationships with salt ions.

Data records where the ionic balance was greater than 25% were removed to ensure that an ion making up a significant component of the salt profile was not missing. In an accurate analysis, the sum of the milliequivalents of major cations and anions will be nearly equal. Bicarbonate and carbonate are two ions that were infrequently measured in the analysis-ready dataset resulting in an expected ionic imbalance to some degree. The dataset was explored for how many data records would be removed based on varying boundary conditions for ionic imbalance and 25% provided a reasonable cut-off combining confidence in data accuracy without excluding a large percentage of data records.

After the above work was completed, the ideal background dataset contained 3,775 data records. The ideal background dataset was then explored to identify definitive patterns present.

3.6 Identifying Background Salinity Patterns

An HCA dendrogram was completed for the ideal background dataset that also included a heat map of the six salt ions (calcium, chloride, magnesium, potassium, sodium, and sulphate). The dendrogram was explored at varying degrees of granularity to determine where relationships between salinity parameters changed meaningfully. An example of a meaningful difference between clusters would be where data records in one cluster were dominated by a strong correlation between magnesium and sulphate while data records in an adjacent cluster were dominated by a strong correlation between calcium and sulphate.

Fifty (50) random data records were selected from each cluster at the chosen level of granularity and Statvis Salt Print Radar plots were generated to visualize similarity of background patterns. Statvis Salinity Radar plots show the five most abundant salinity ions (calcium, chloride, magnesium, sodium, and sulphate) on the points of a pentagon. The scale used on the radar plots is percent contribution. Potassium is not included as it typically has too low a contribution to provide a diagnostic visual pattern. Fifty (50) data records were chosen as it is a large enough number to show the range of data records present but small enough to review and understand visually. These data records are expected to represent the cluster as they were chosen randomly. This was done to provide a visual representation of the range of patterns in the a priori clusters but not to assign data records to clusters. From the 50 random data records, an average radar plot was generated. The average radar plot provides a visualization of the salt pattern representative of the background cluster.

3.7 Applying Background Patterns to the Full Pilot Area Dataset

Boundary conditions for the six salt ions (calcium, chloride, magnesium, potassium, sodium, and sulphate) and EC were defined for each background cluster in the ideal background dataset. Boundary conditions were defined using the minimum and maximum percent contribution of each of the six salt ions and EC value in each background cluster from the ideal background dataset. To be designated as belonging to a background cluster, a data record had to have measured values of all six salt ions and EC within the minimum and maximum boundary conditions set for a particular cluster. The master dataset of 23,821 records was compared to these boundary conditions and records that did not fit into one or more background cluster were removed from the final dataset as they were considered to show anthropogenic influence.

4. Results

The HCA dendrogram and heatmap for calcium, chloride, potassium, sodium, magnesium, sulphate, and EC generated from the ideal background dataset are shown in Figure 3. There are five main HCA clusters—coloured red, green, blue, brown, and teal respectively—as well as 12 more granular clusters identified. The 12 clusters represent different background patterns present in the pilot area.

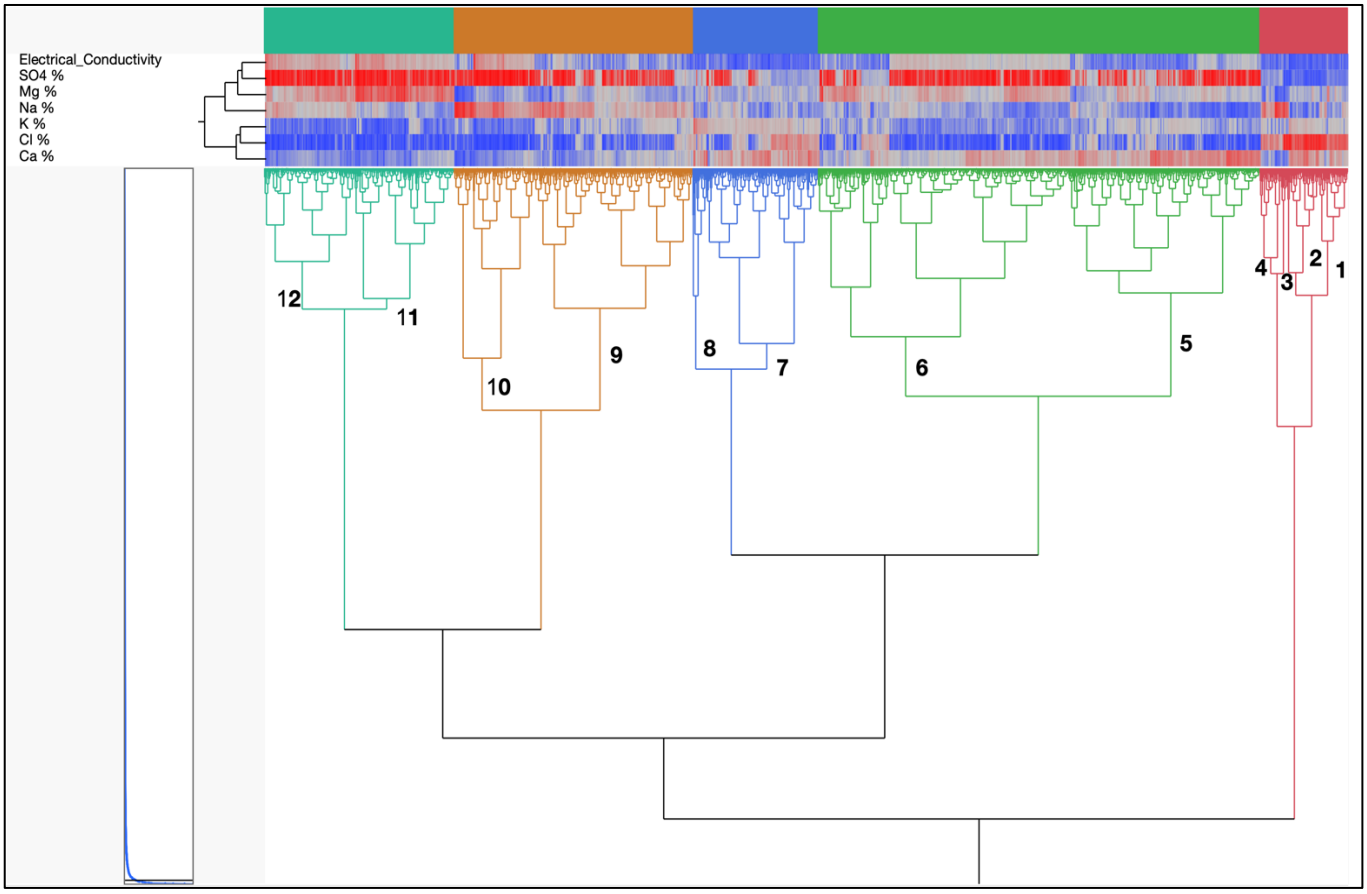


Figure 3. HCA with five main coloured HCA clusters and 12 descriptive clusters representing background fingerprints in the pilot area.

Statvis Salt Prints were generated for 50 random data records from each of the 12 background patterns identified. Radar plots and summary statistics for each background cluster are included as Appendix A to this report.

Based on the 12 background patterns identified in the 3,775 data records included in the ideal background dataset, 18,105 data records of the 23,821 data records in the entire pilot area dataset were identified as representative of background.

5. Conclusions

Measured values of salinity parameters in background may exceed regulatory guidelines. The objectives of the current scope of work defined for Statvis in Phase 1 of the ABSQS were to harmonize and clean the data, as well as develop a workflow to remove impacted data records leaving only background data records in the ABSQS database. These objectives were achieved within the defined scope of work.

Throughout the process of collecting data from data providers several challenges were identified. Important metadata items—including units of measure and analytical methods used—were sometimes not provided. This lengthened the time required for the data collection phase. Collecting coordinates for each individual data point leads to increased predictive power in analysis of patterns and trends in soil chemistry datasets.

The nine salinity parameters chosen for the ABSQS (pH, EC, SAR, chloride, sulphate, calcium, magnesium, potassium, and sodium) are the most regularly reported across the salinity analytical packages found in the provided datasets.

The salinity analysis workflow derived to identify background data records, so that impacted data records could be removed from the final ABSQS database, provided stable and replicable results. Of the 23,821 data records in the entire pilot area dataset, 18,105 data records were identified as background.

6. Recommendations

Based on the results of the salinity data analysis, the following are recommended:

- The creation of a prototype of the ABSQS for a pilot area (Phase 1) was successful and the prototype should be shared with stakeholders for testing and feedback.
- The ABSQS should be expanded to the full provincial scale (i.e., carry on to Phases 2 and 3).
- The nine salinity parameters used for the pilot area (pH, EC, SAR, chloride, sulphate, calcium, magnesium, potassium, and sodium) should be carried forward to future phases of the project.
- As a significant amount of the data provided for the ABSQS came directly from analytical laboratories, an attempt should be made to add geospatial coordinates to the list of metadata items included in lab databases going forward.
- Future phases or similar projects should allot additional time or the data collection phase.
- Although a clear template for data formatting was provided, future phases or similar projects should provide additional guidance on data and metadata requirements (i.e., must-haves vs. nice-to-haves).
- Datasets with geospatial coordinates for individual data records should continue to be solicited opportunistically.

7. Statement of Limitations

This report was prepared for the exclusive use of the client identified herein. The report, which specifically includes all tables, figures, and appendices, is based on data and information collected or provided during the work conducted by Statvis Analytics Inc. and is based solely on the conditions of the site and data obtained by Statvis Analytics Inc. as described in this report. Information and data provided to Statvis Analytics Inc. has not been independently verified.

The services performed as described in this report were conducted in a manner consistent with the level of care and skill normally exercised by other environmental professionals currently practicing under similar conditions.

Any use a third party makes of this report, or any reliance on or decisions to be made based on it, are the responsibilities of such third parties. Statvis Analytics Inc. accepts no responsibility for damages, if any, suffered by any third party as a result of decisions made or actions based on this report.

The content of this report is based on data and information collected or provided during our assessment, our present understanding of site conditions and our professional judgement in light of such information at the time of this report. This report provides a professional opinion and therefore no warranty is expressed, implied, or made as to the conclusions and recommendations offered in this report. This report does not provide a legal opinion regarding compliance with applicable laws. It should be noted that regulatory statutes and the interpretation of regulatory statutes are subject to change. The findings and conclusions of this report are valid only as of the date of this report. If new information is discovered in future work, Statvis Analytics Inc. should be engaged to re-evaluate the conclusions of this report and provide amendments as required.

8. Closure

We trust that the information presented in this report meets your current requirements. Should you have any questions or require additional information, please do not hesitate to contact the undersigned.

Sincerely,

Statvis Analytics Inc.

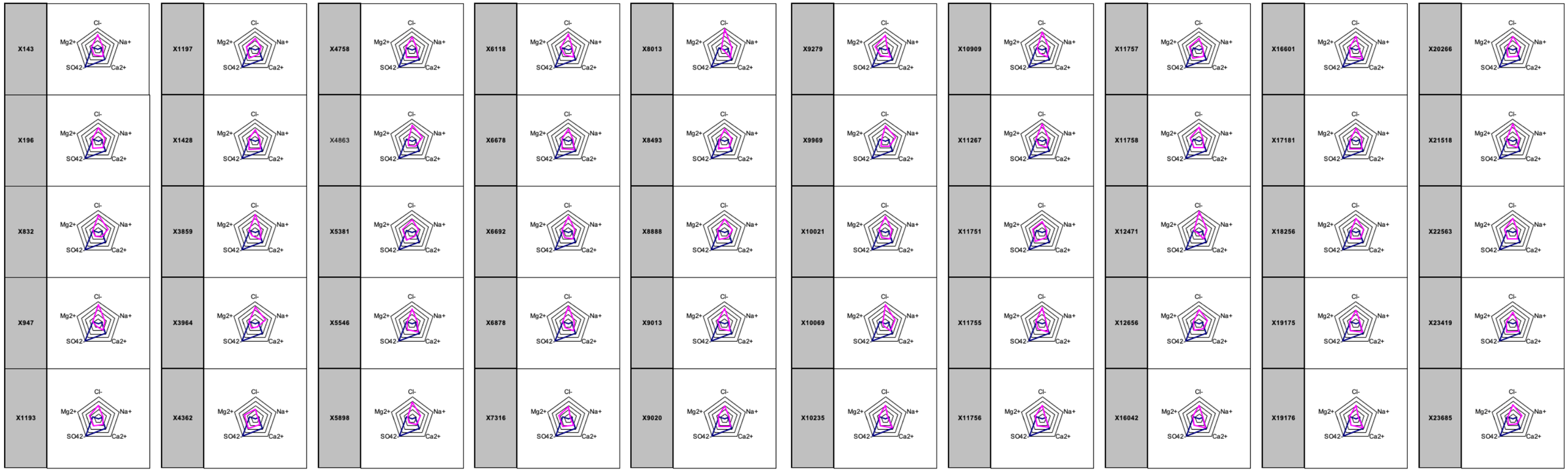
Paul Fuellbrandt, P.Ag., PMP
Principal Environmental Scientist

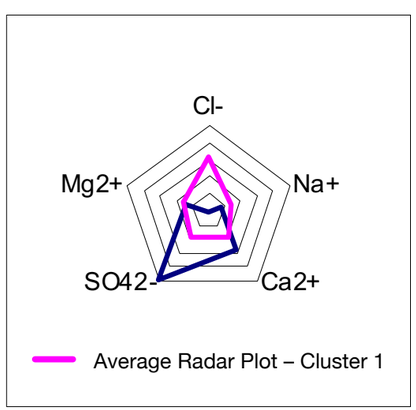
Court Sandau, PhD
Principal Chemist

statvis Trust Through Data

Appendix A. Radar Plots and Summary Statistics for Background Clusters

Cluster 1– 50 Randomly Selected Radar Plots



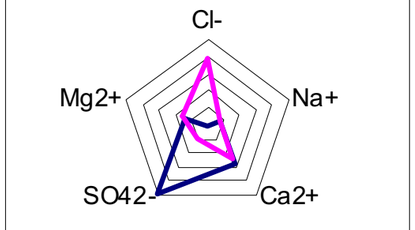


Summary Statistics from Background Dataset that Define Cluster 1

	K %	Mg %	Ca %	Na %	SO4 %	Cl %
Mean	2.43	30.79	39.18	27.61	35.35	64.65
Min	0.92	10.66	19.21	10.72	1.33	37.92
Max	4.30	52.43	57.68	50.53	62.08	98.67
Median	2.40	29.21	38.73	26.27	37.41	62.59
Quantiles95	3.54	47.61	52.57	46.90	51.04	89.49

Cluster 2 – 50 Randomly Selected Radar Plots

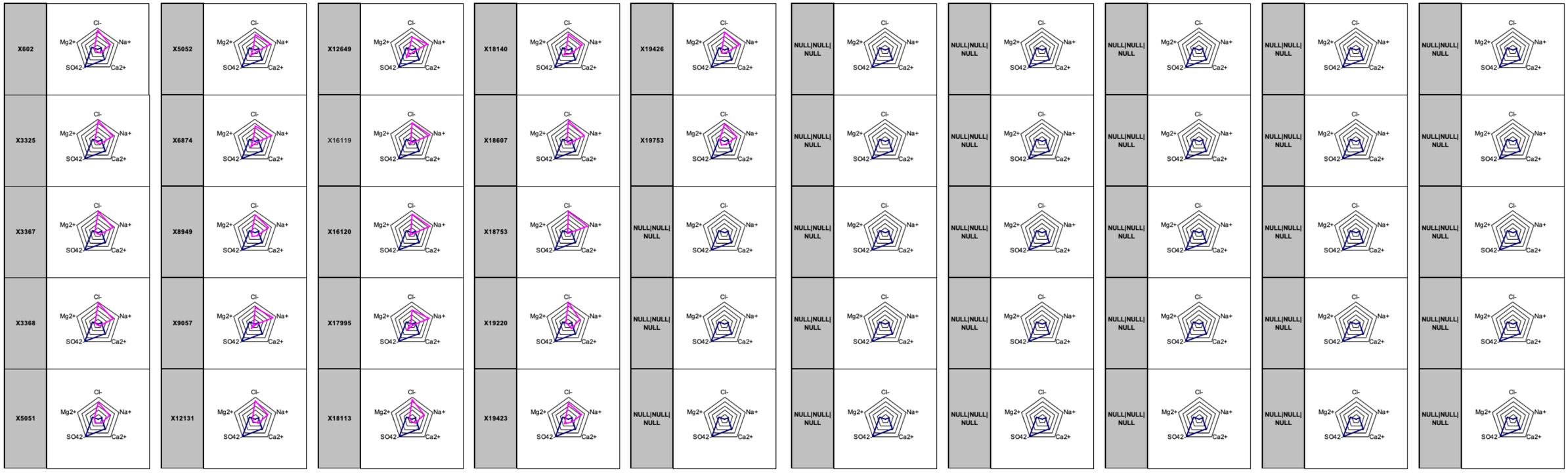
X107	X2881	X5008	X10081	X11851	X12612	X17018	X17367	X18444	X19428
X1104	X2939	X5542	X10190	X11858	X13107	X17092	X17369	X18522	X19618
X2002	X4137	X6905	X11741	X11891	X16046	X17103	X17522	X18959	X20608
X2132	X4732	X8371	X11842	X11893	X16096	X17109	X17683	X19085	X21530
X2399	X4886	X10015	X11843	X12168	X16216	X17316	X17760	X19086	X23585

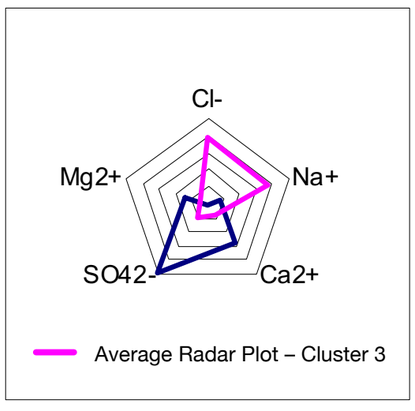


Summary Statistics from Background Dataset that Define Cluster 2

	K %	Mg %	Ca %	Na %	SO4 %	Cl %
Mean	4.68	31.07	50.01	14.25	19.80	80.20
Min	1.86	12.53	31.32	3.93	2.82	54.44
Max	13.75	47.86	70.82	32.11	45.56	97.18
Median	4.02	31.08	49.88	13.60	18.54	81.46
Quantiles95	9.79	41.18	59.66	24.90	39.86	95.94

Cluster 3 – 50 Randomly Selected Radar Plots

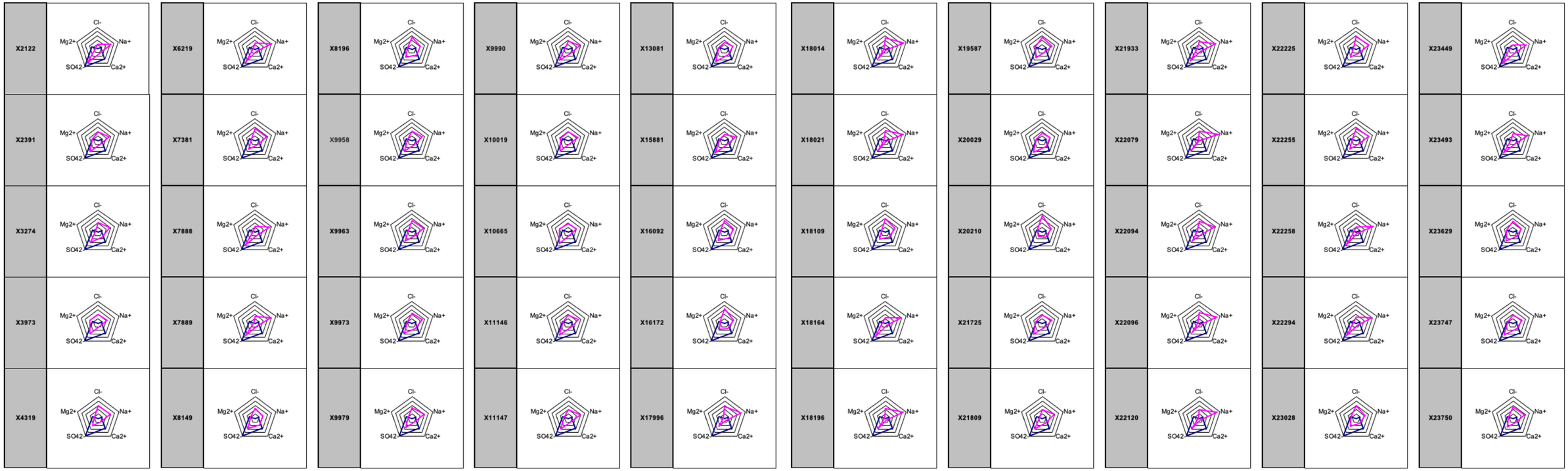


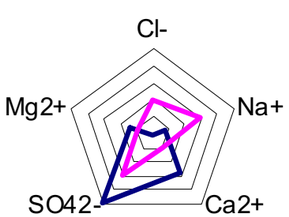


Summary Statistics from Background Dataset that Define Cluster 3

	K %	Mg %	Ca %	Na %	SO4 %	Cl %
Mean	1.88	10.51	15.38	72.23	19.45	80.55
Min	0.80	2.06	4.68	56.51	2.21	56.48
Max	4.79	22.04	28.49	92.47	43.52	97.79
Median	1.66	11.40	14.45	71.67	18.23	81.77
Quantiles95	4.70	21.33	27.92	91.63	42.42	97.37

Cluster 4 – 50 Randomly Selected Radar Plots

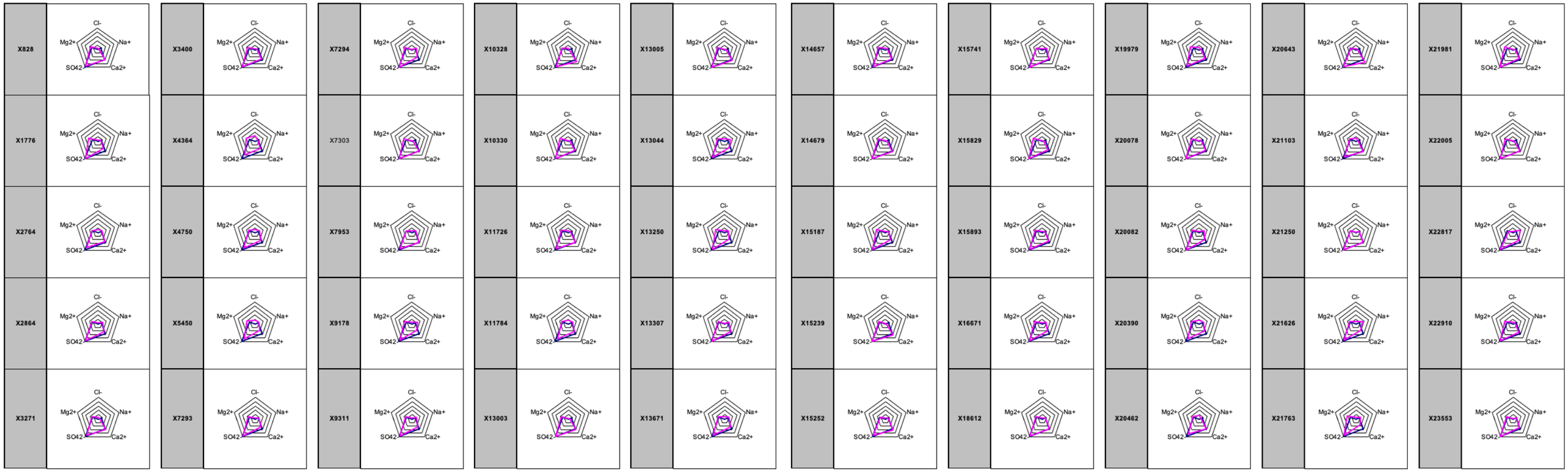


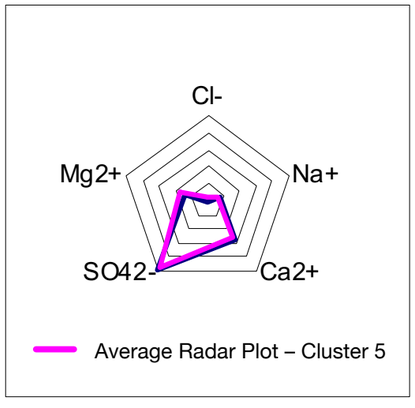


Summary Statistics from Background Dataset that Define Cluster 4

	K %	Mg %	Ca %	Na %	SO4 %	Cl %
Mean	2.38	17.11	21.26	59.24	58.09	41.91
Min	0.45	1.54	2.63	36.51	25.64	20.10
Max	6.78	39.77	38.31	94.14	79.90	74.36
Median	2.10	17.80	22.27	55.16	58.09	41.91
Quantiles95	6.03	34.05	35.17	87.48	76.17	57.35

Cluster 5 – 50 Randomly Selected Radar Plots

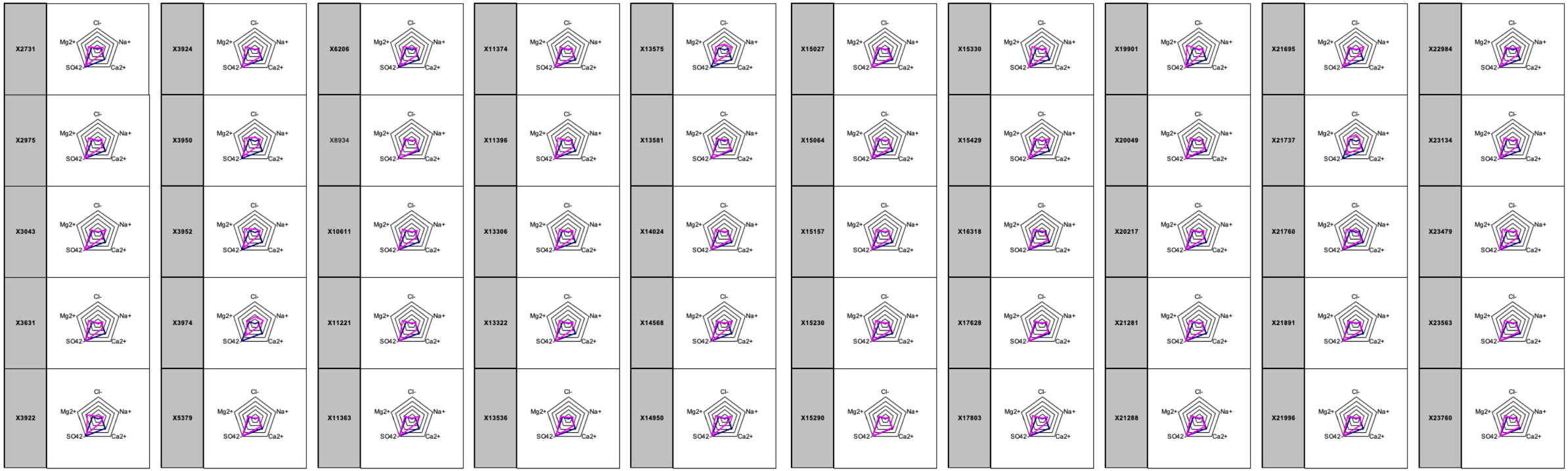


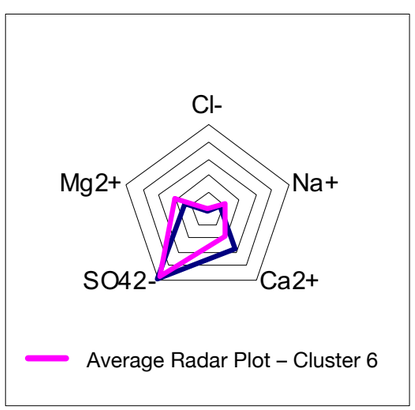


Summary Statistics from Background Dataset that Define Cluster 5

	K %	Mg %	Ca %	Na %	SO4 %	Cl %
Mean	1.95	33.02	49.43	15.61	93.49	6.51
Min	0.35	3.20	27.77	1.34	74.83	0.17
Max	4.50	50.47	83.03	43.92	99.83	25.17
Median	1.77	33.39	49.42	13.95	94.79	5.21
Quantiles95	3.66	42.68	62.05	31.88	99.49	16.86

Cluster 6 – 50 Randomly Selected Radar Plots

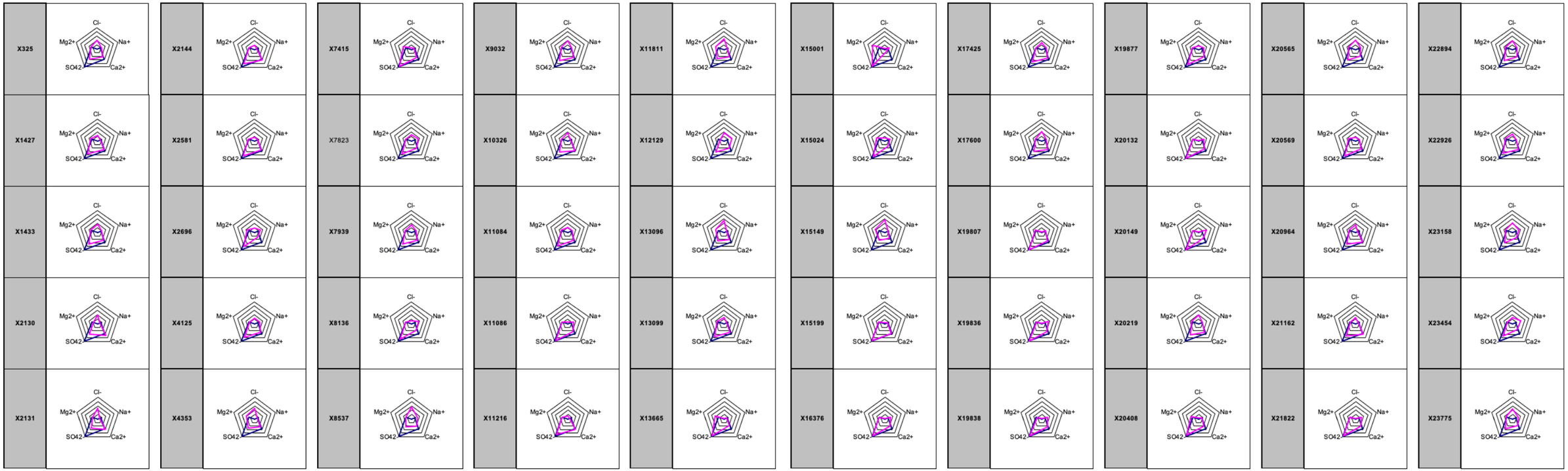


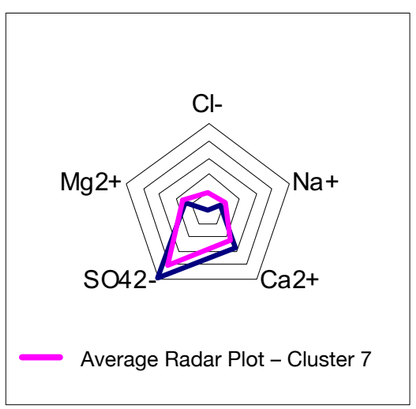


Summary Statistics from Background Dataset that Define Cluster 6

	K %	Mg %	Ca %	Na %	SO4 %	Cl %
Mean	1.24	41.41	36.91	20.43	94.48	5.52
Min	0.13	27.17	10.96	3.72	57.58	0.12
Max	3.46	73.50	56.41	41.31	99.88	42.42
Median	1.18	41.25	37.03	20.33	97.91	2.09
Quantiles95	2.14	51.99	49.34	33.67	99.67	27.25

Cluster 7 – 50 Randomly Selected Radar Plots

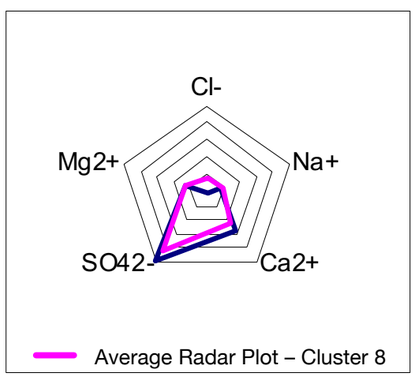




Summary Statistics from Background Dataset that Define Cluster 7

	K %	Mg %	Ca %	Na %	SO4 %	Cl %
Mean	4.05	29.25	45.06	21.64	76.74	23.26
Min	0.87	5.71	12.17	2.87	38.82	0.82
Max	8.67	58.73	79.77	46.82	99.18	61.18
Median	3.93	28.78	44.64	21.78	78.68	21.32
Quantiles95	6.25	39.26	59.89	35.84	98.17	49.90

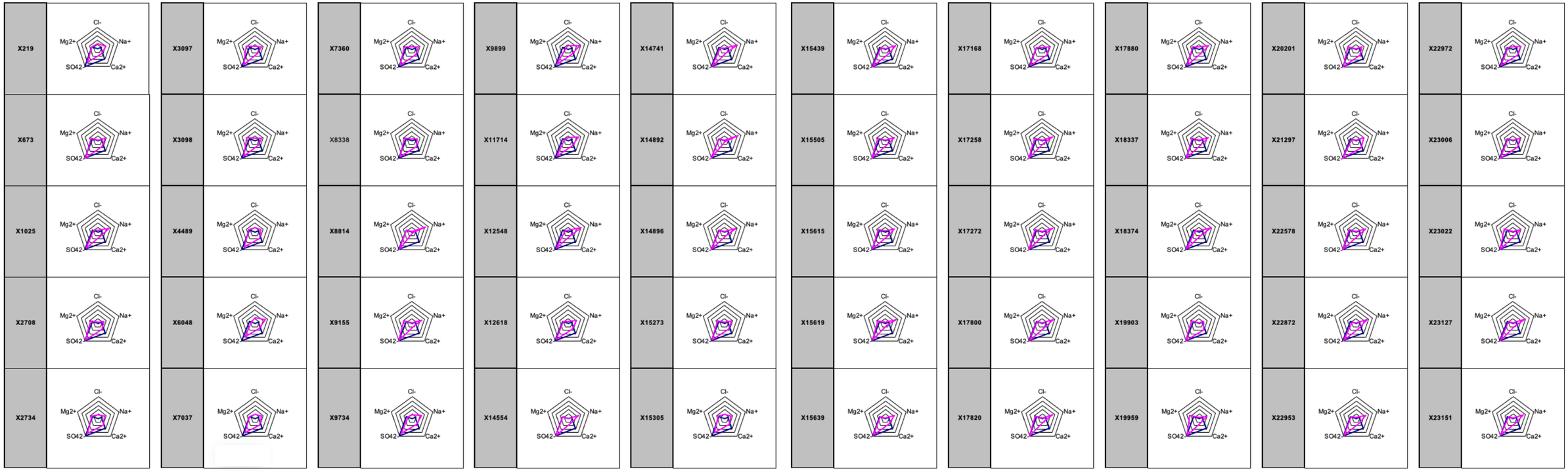
Cluster 8 – 50 Randomly Selected Radar Plots

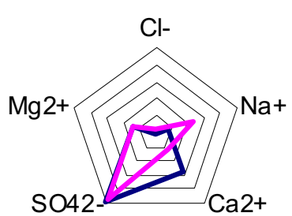


Summary Statistics from Background Dataset that Define Cluster 8

	K %	Mg %	Ca %	Na %	SO4 %	Cl %
Mean	10.19	26.42	44.31	19.08	81.68	18.32
Min	5.79	4.91	18.12	5.39	51.43	2.01
Max	30.11	35.77	68.82	46.04	97.99	48.57
Median	8.86	28.33	43.72	17.90	84.08	15.92
Quantiles95	20.22	34.24	67.66	36.71	96.12	42.42

Cluster 9 – 50 Randomly Selected Radar Plots

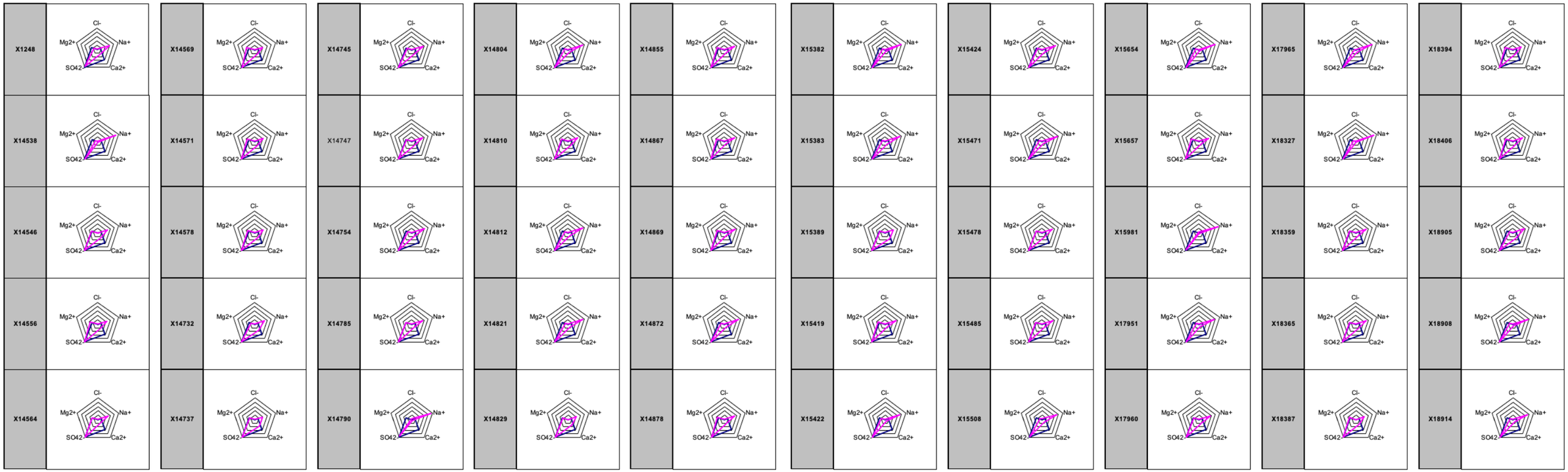


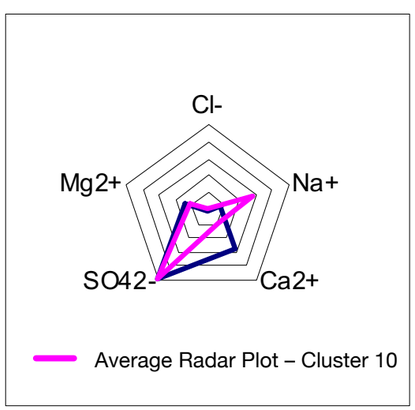


Summary Statistics from Background Dataset that Define Cluster 9

	K %	Mg %	Ca %	Na %	SO4 %	Cl %
Mean	2.03	27.51	26.29	44.17	93.55	6.45
Min	0.19	3.67	3.91	22.48	69.80	0.14
Max	4.96	45.46	41.11	76.16	99.86	30.20
Median	1.89	27.73	27.38	43.02	96.02	3.98
Quantiles95	3.82	39.32	37.43	63.66	99.46	20.18

Cluster 10 – 50 Randomly Selected Radar Plots



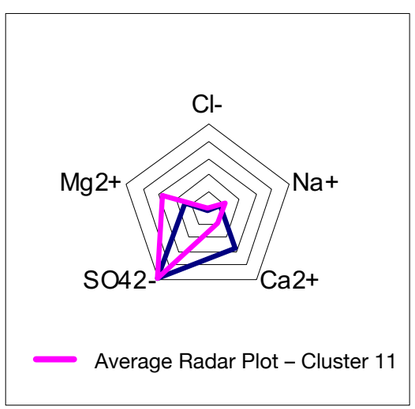


Summary Statistics from Background Dataset that Define Cluster 10

	K %	Mg %	Ca %	Na %	SO4 %	Cl %
Mean	0.95	20.54	18.60	59.91	98.69	1.31
Min	0.17	0.71	1.30	36.22	90.04	0.09
Max	2.73	37.25	43.70	97.32	99.91	9.96
Median	0.89	21.32	18.87	54.42	99.27	0.73
Quantiles95	1.98	33.92	34.34	89.09	99.87	5.67

Cluster 11 – 50 Randomly Selected Radar Plots

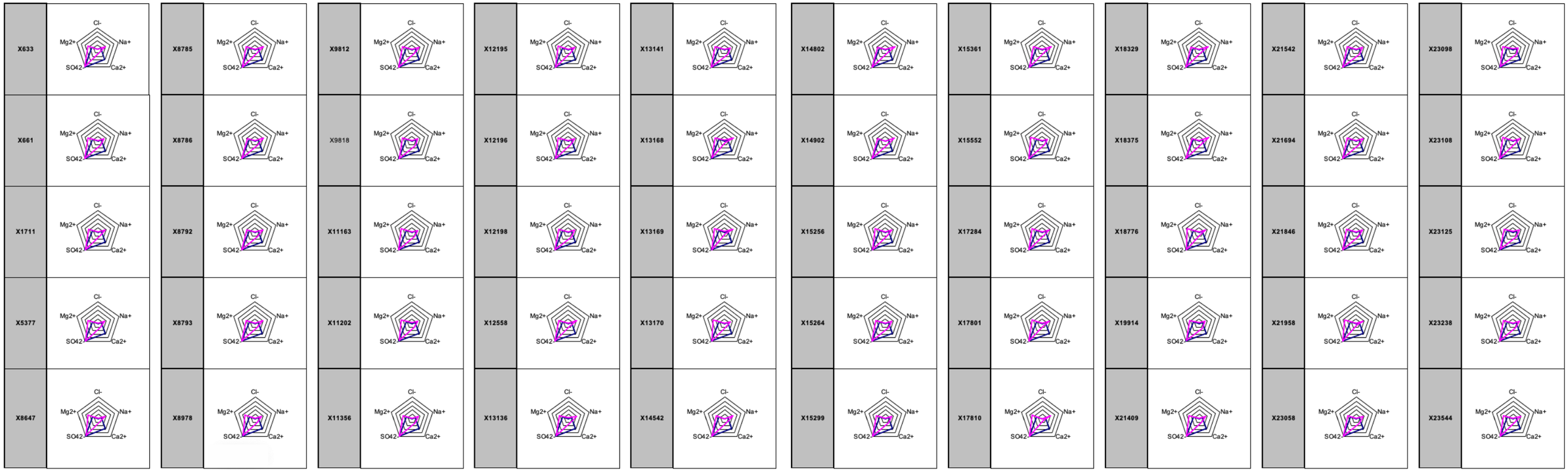
X719		X7784		X10333		X15004		X15154		X15366		X18660		X20489		X21883		X22040	
X1352		X8054		X10429		X15007		X15156		X16676		X19829		X21404		X21900		X23530	
X3037		X10041		X11159		X15013		X15162		X17813		X20024		X21777		X21947		X23548	
X4535		X10044		X13171		X15017		X15165		X17984		X20108		X21843		X21949		X23551	
X7345		X10332		X14995		X15018		X15228		X18651		X20314		X21877		X21969		X23805	

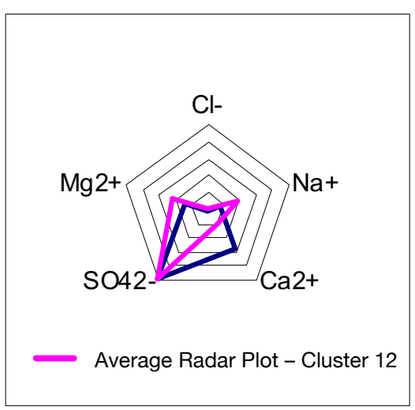


Summary Statistics from Background Dataset that Define Cluster 11

	K %	Mg %	Ca %	Na %	SO4 %	Cl %
Mean	0.97	55.89	21.11	22.03	98.57	1.43
Min	0.18	41.50	5.62	5.08	92.07	0.08
Max	5.03	70.39	39.57	38.57	99.92	7.93
Median	0.73	55.93	20.85	21.26	99.19	0.81
Quantiles95	2.71	65.52	33.77	34.62	99.81	4.47

Cluster 12 – 50 Randomly Selected Radar Plots





Summary Statistics from Background Dataset that Define Cluster 12

	K %	Mg %	Ca %	Na %	SO4 %	Cl %
Mean	0.76	43.88	19.96	35.40	98.19	1.81
Min	0.17	29.95	4.95	21.25	87.74	0.07
Max	2.01	58.06	30.15	58.60	99.93	12.26
Median	0.73	44.05	20.25	33.78	98.87	1.13
Quantiles95	1.23	52.50	27.36	47.92	99.87	5.41