



Multivariate Statistical Analysis in Water Quality

Karen R. Ryberg
U.S. Geological Survey

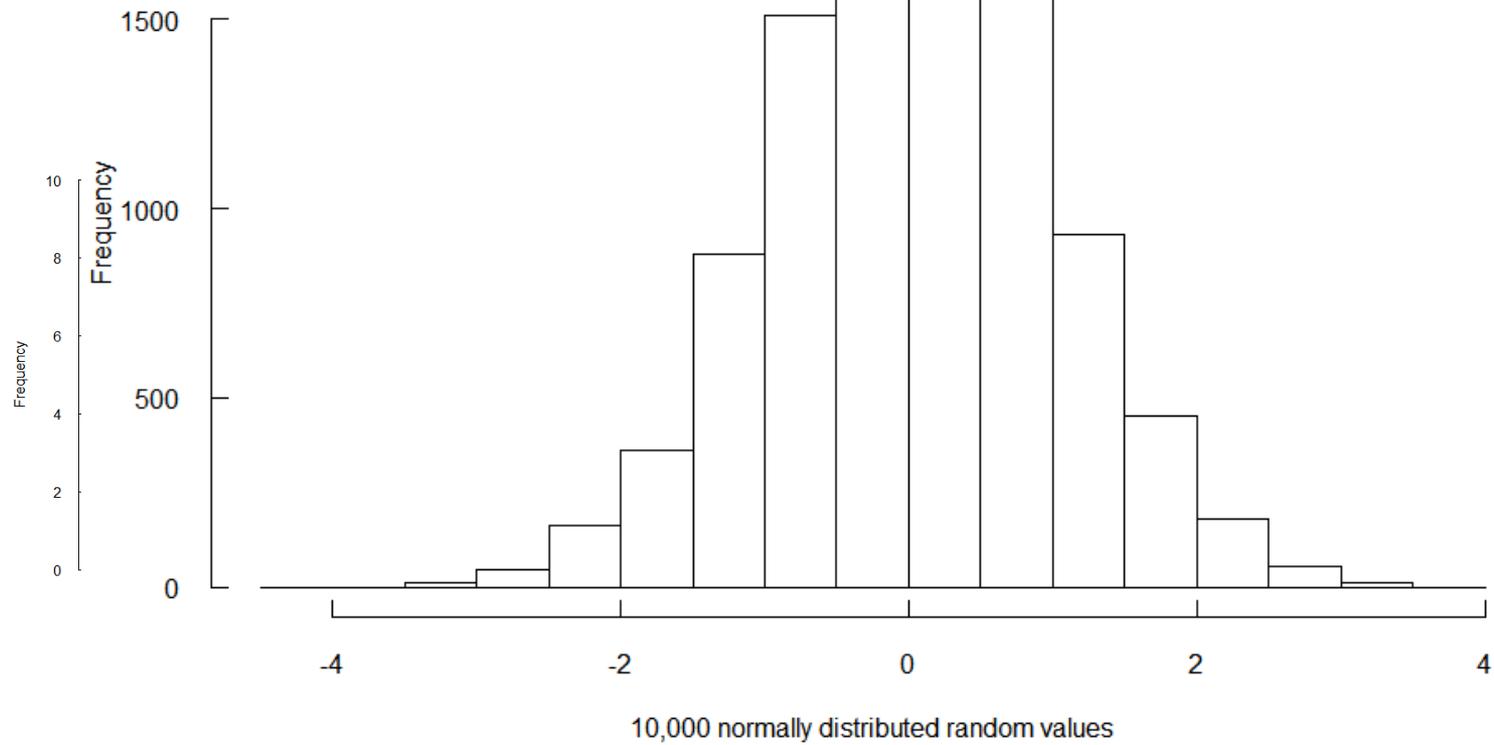
National Water Quality Monitoring Council Webinar Series
November 29, 2017

Exploratory Graphical and Numerical Analysis

- Histogram
- Plots: X, Y, and Z and pairs plots
- Correlation

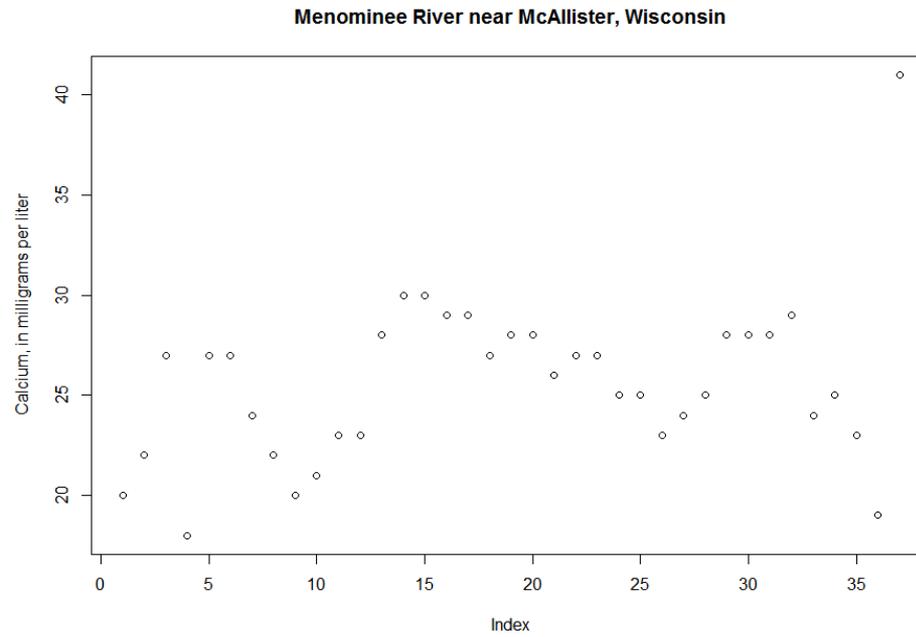


Hist



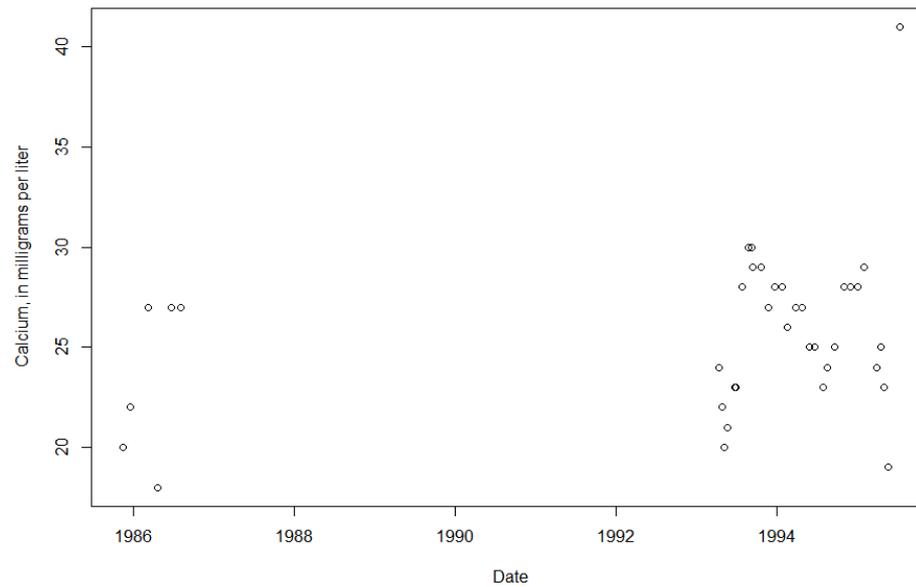


X Plot



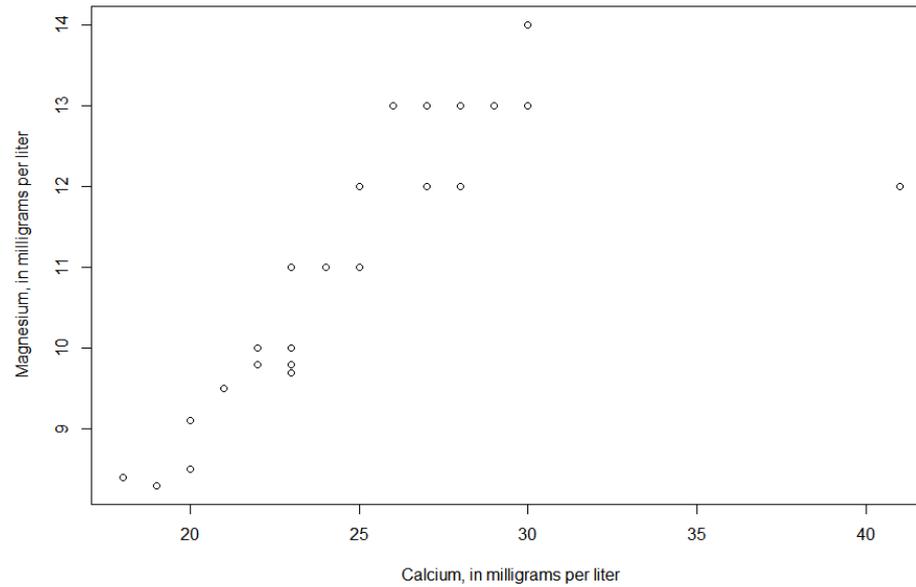


X-Y Plot



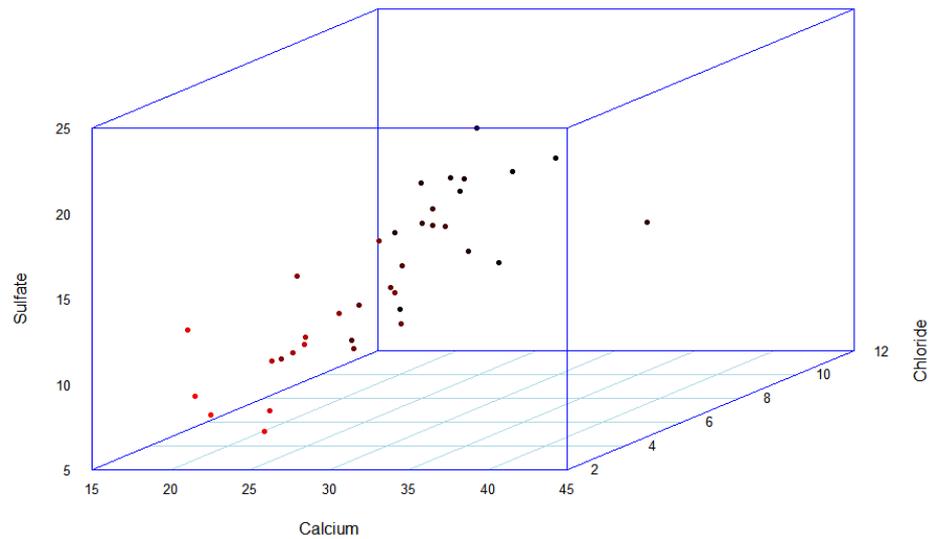


X-Y Plot

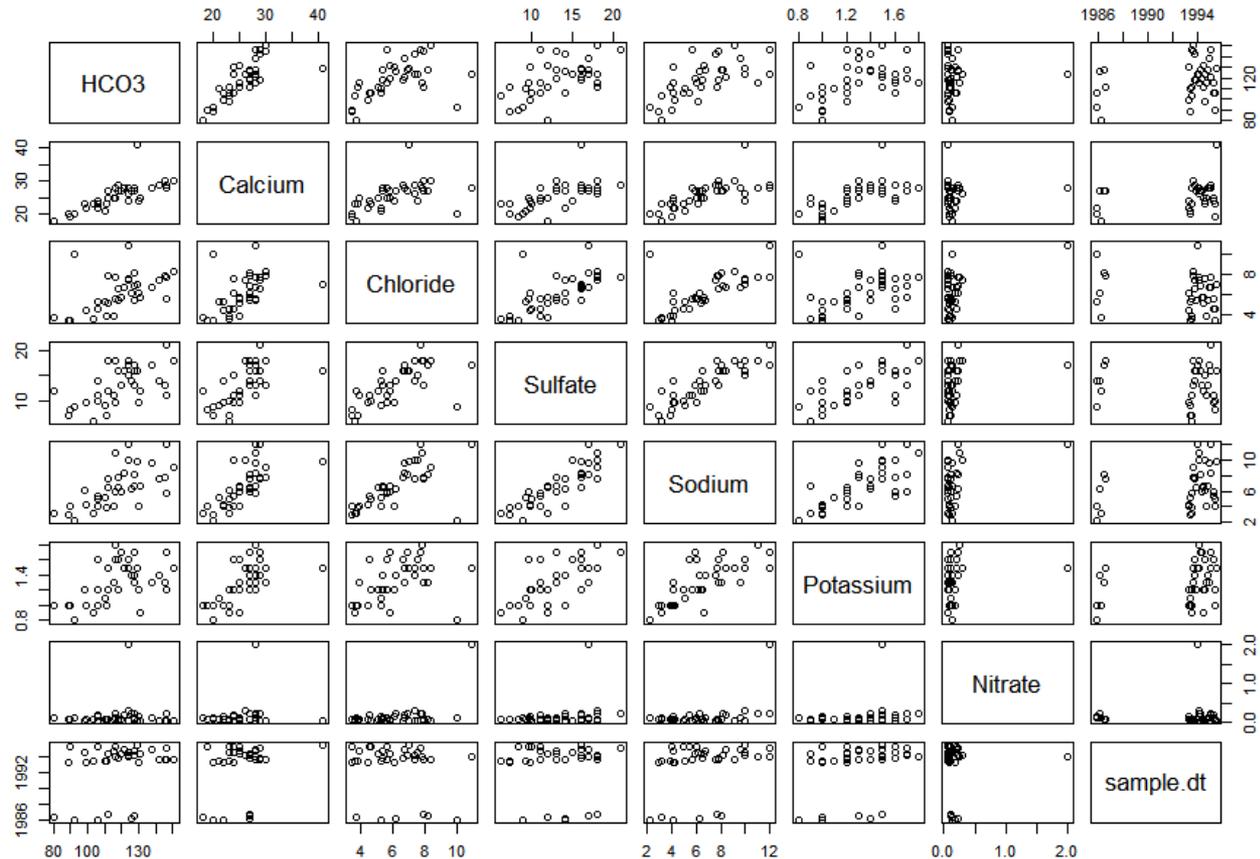


X-Y-Z Plot

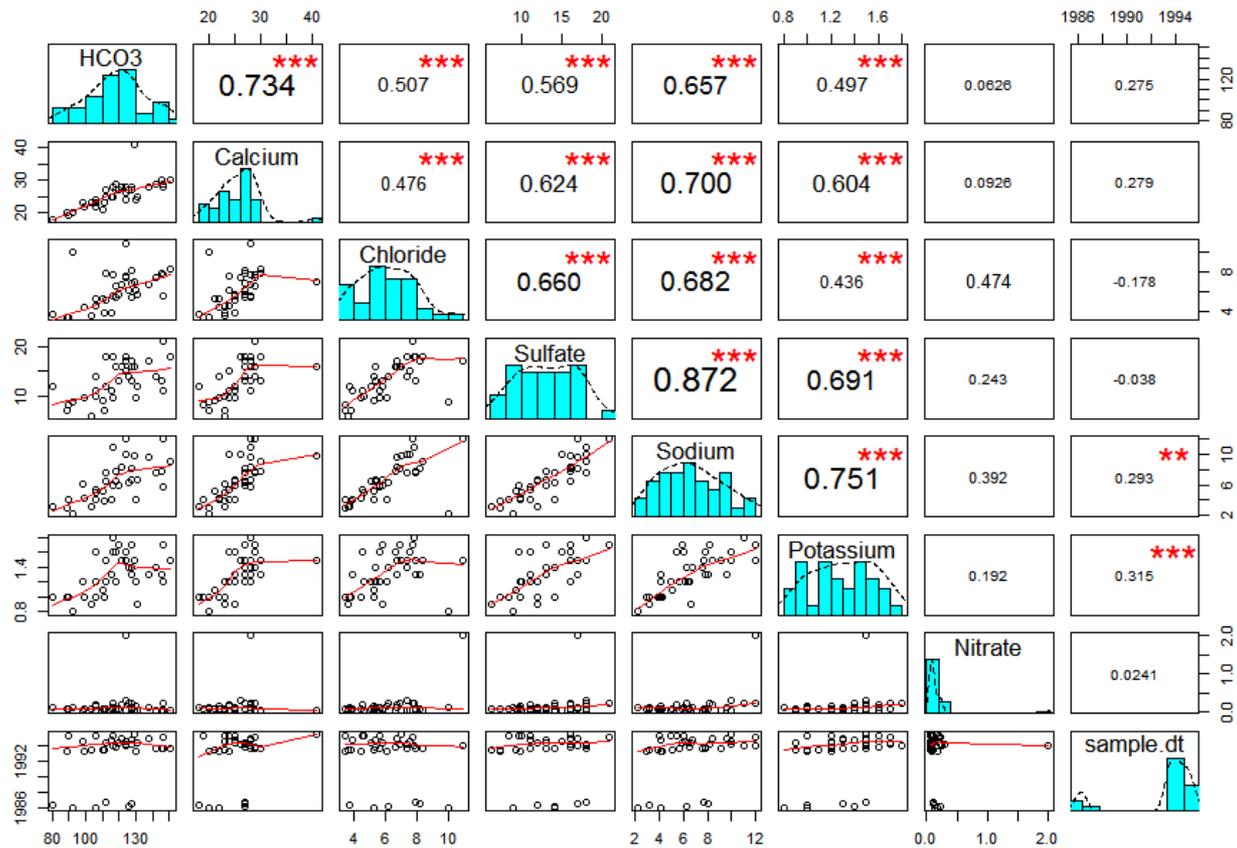
Menominee River near McAllister, Wisconsin



Pairs Plots



Pairs Plots



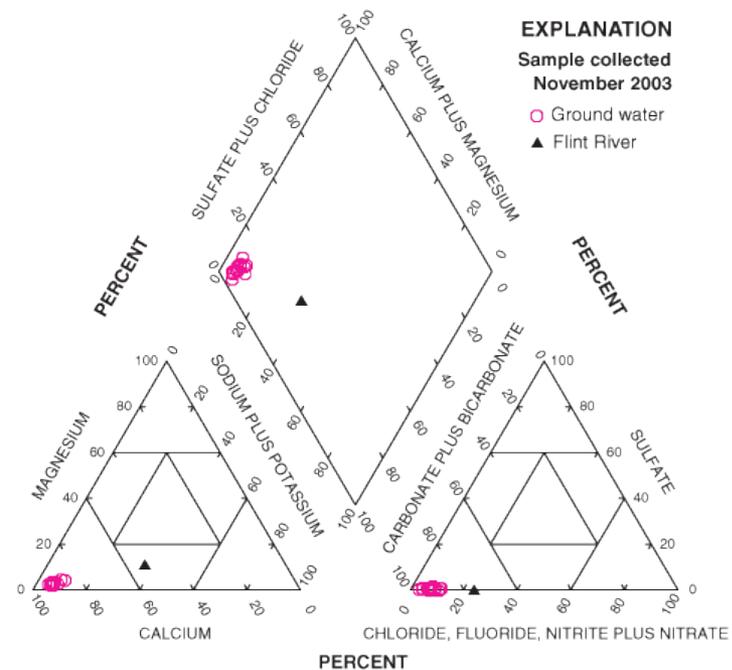
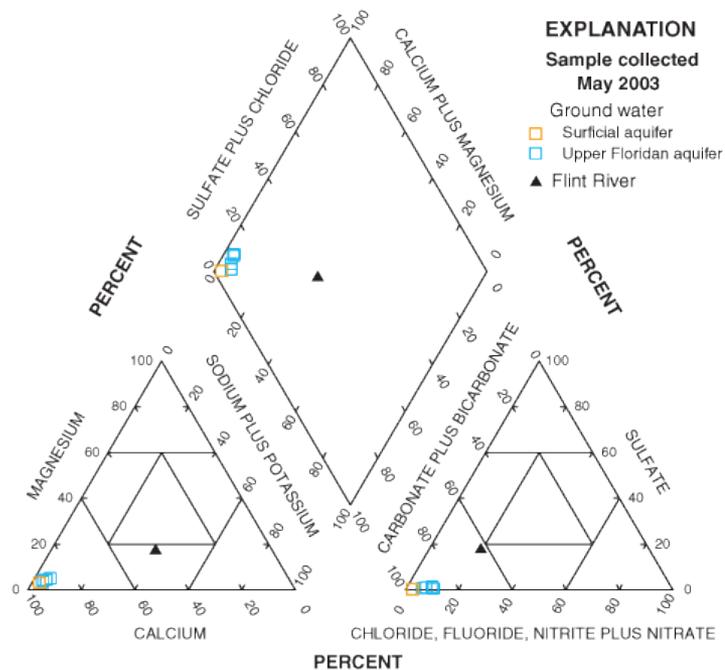


Notes on Correlation

- Pearson's r (Pearson correlation coefficient, Pearson product-moment correlation coefficient)
 - Assumes a linear relation
 - Often what people are referring to when they say correlation, but not always appropriate
- Kendall's tau correlation (Kendall's rank correlation coefficient)
 - Non-parametric, based on ranks, when one value is large the other value tends to be larger, but not necessarily a linear relation, a measure of ordinal association
- Spearman's rho (Spearman's rank correlation coefficient)
 - Non-parametric, based on ranks and the assumption that two variables follow a monotonic function

Piper/Ternary/Trilinear Diagram

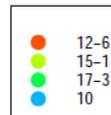
Leeth and others, 2005



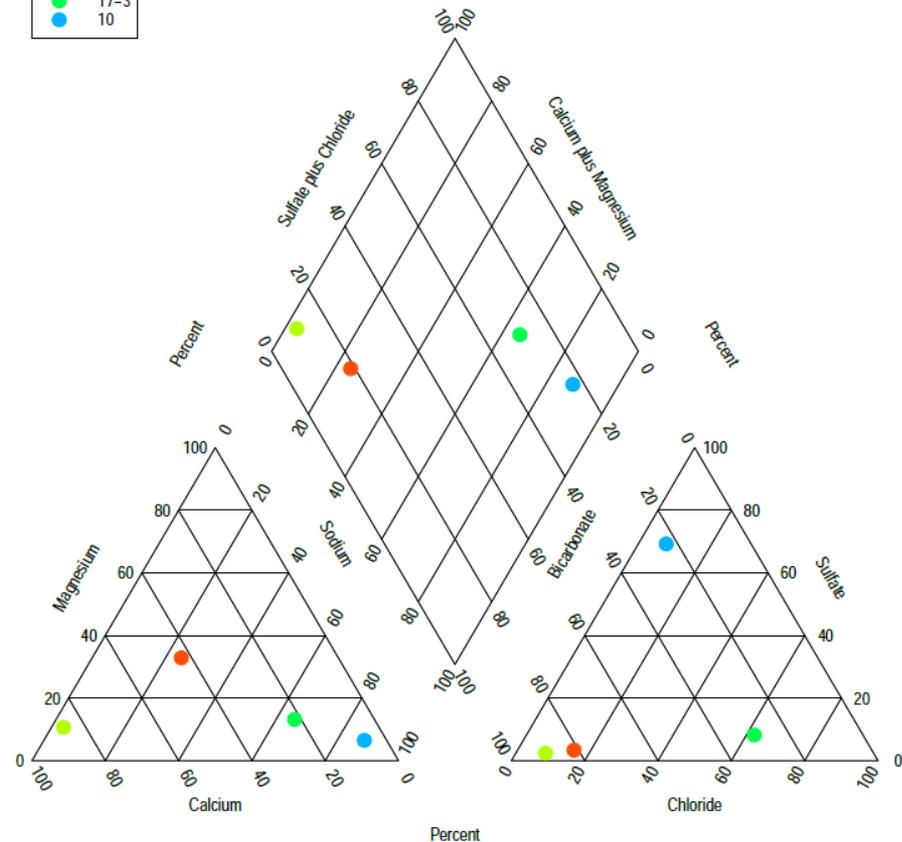
Rectangular version for Python and ArcGIS -
<http://python.hydrology-amsterdam.nl/>

Trilinear version for R <https://cran.r-project.org/web/packages/hydrogeo/>

Piper/Ternary/Trilinear Diagram



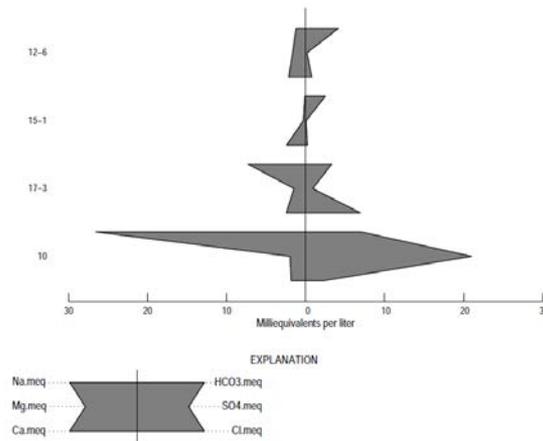
Uses USGS R packages
smwrGraphs and smwrData
(citation on next slide)



Stiff Diagram

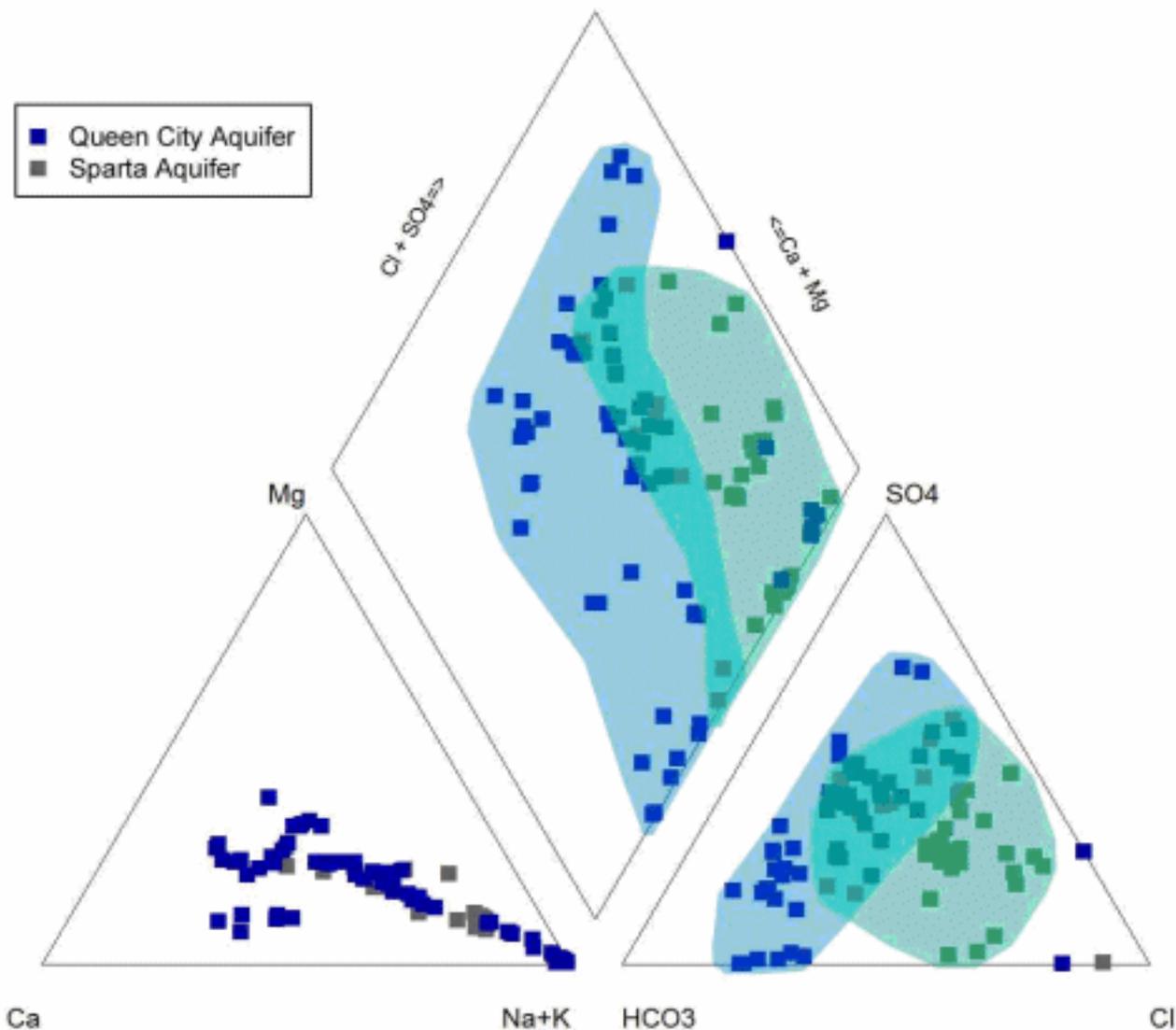
- Similar to Piper, but less detailed could be placed graphically on a map

Uses USGS R packages
smwrGraphs and smwrData



Lorenz, D.L., and Diekoff, A.L., 2017, smwrGraphs—An R package for graphing hydrologic data, version 1.1.2: U.S. Geological Survey Open-File Report 2016–1188, 17 p., <https://doi.org/10.3133/ofr20161188>.
Lorenz, D.L., 2015, smwrData—An R package of example hydrologic data, version 1.1.1: U.S. Geological Survey Open-File Report 2015–1103, 5 p., <http://dx.doi.org/10.3133/ofr20151103>.

Stiff Diagrams on a Map



2006 Progress Report: E Groundwater Chemistry County, Texas and Techn Educational Assistance t Conservation Districts ir <https://cfpub.epa.gov/n dex.cfm/fuseaction/disq tract/7571/report/2006>

Stiff Diagrams on a Map

Bartos, T.T. and Muller Ogle, K., 2002, Water quality and environmental isotopic analyses of ground-water samples collected from the Wasatch and Fort Union formations in areas of coalbed methane development—Implications to recharge and ground-water flow, eastern Powder River Basin, Wyoming: U.S. Geological Survey Water-Resources Investigations Report 02-4045, 88 p., <https://pubs.usgs.gov/wri/wri024045/>.

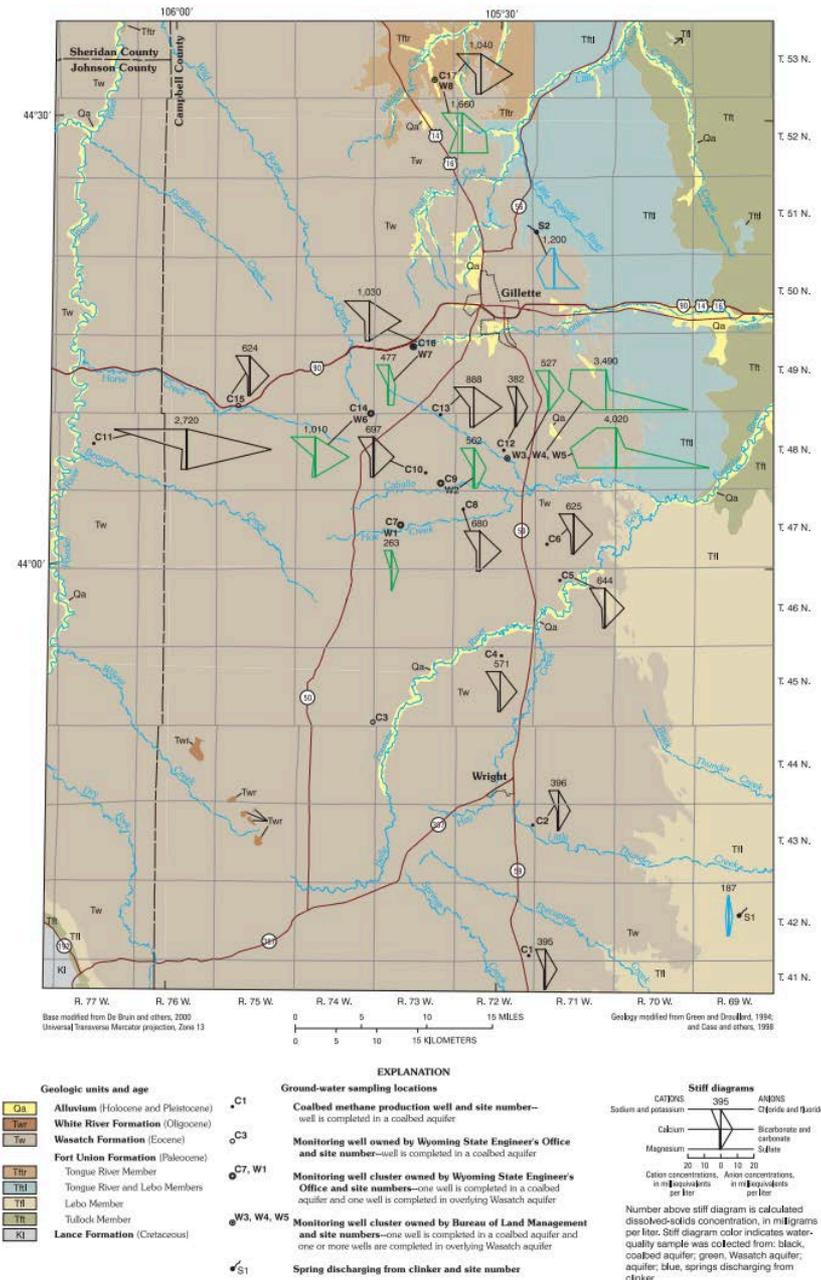


Figure 14. Stiff diagrams for ground-water samples collected from coalbed methane production wells, monitoring wells, and springs, and generalized geology in the study area, eastern Powder River Basin, Wyoming, 1999. Bicarbonate and dissolved-solids concentrations are estimated for well W7.



Outlier Detection

- Pairs plots
- Plots against time
- Check cation/anion balance
 - The ion equivalents of the cations should approximately match that of the anions
 - Lab should check this, often can be a problem in legacy data though
 - Example calculator here <http://users.tinyonline.co.uk/chrisshort/ib.htm>



Multiple Regression

- Many underlying assumptions that need to be checked for model to be appropriate
- Statistical Methods in Water Resources (Helsel and Hirsch, 1992), excellent resource for this.
- New version coming in 2018, with examples in R

Assumptions Necessary for OLS Purposes

Assumption	Purpose			
	Predict y given x	Predict y and a variance for the prediction	Obtain best linear unbiased estimator of y	Test hypotheses, estimate confidence or prediction intervals
(1) Model form is correct: y is linearly related to x	+	+	+	+
(2) Data used to fit the model are representative of data of interest.	+	+	+	+
(3) Variance of the residuals is constant (is homoscedastic). It does not depend on x or on anything else (e.g. time).		+	+	+
(4) The residuals are independent.			+	+
(5) The residuals are normally distributed.				+

Page 225 of
Helsel, D.R. and R. M. Hirsch, 2002,
Statistical Methods in Water Resources:
U.S. Geological Survey Techniques of Water
Resources Investigations, Book 4, chapter
A3, 522 p.,
<https://pubs.usgs.gov/twri/twri4a3/>.



Table 9.1 Assumptions necessary for the purposes to which OLS is put.

+: the assumption is required for that purpose.

Regression Diagnostics

- Excellent Regression Diagnostics in
 - Base R—includes the residual plots—very important to study plots
 - car package—Companion to Applied Regression; functions and datasets to accompany J. Fox and S. Weisberg, *An R Companion to Applied Regression, Second Edition*, Sage, 2011.
 - rms package—Regression modeling strategies by Frank Harrell regression modeling, testing, estimation, validation, graphics, prediction

R Core Team, 2017, R: A language and environment for statistical computing: R Foundation for Statistical Computing, Vienna, <https://www.r-project.org/>.

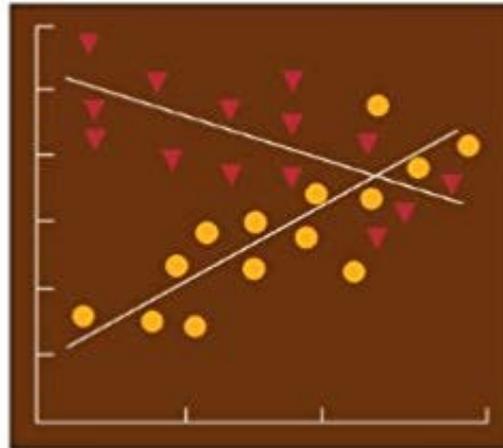
Fox, J. and Weisberg, S., 2011, *An R companion to applied regression, (2nd ed.)*: Thousand Oaks, Calif., Sage, <http://socserv.socsci.mcmaster.ca/jfox/Books/Companion>

Harrell, Jr., F.E., 2016, rms—Regression Modeling Strategies, R package version 5.0-1, <https://CRAN.R-project.org/package=rms>

General Text for Linear Regression and ANOVA

APPLIED LINEAR STATISTICAL MODELS

FIFTH EDITION



Kutner

Nachtsheim

Neter



Cluster Analysis

- Many versions of cluster analysis
- Güler and others (2002) described hierarchical cluster analysis as “an efficient means to recognize groups of samples that have similar chemical and physical characteristics.”
- Hierarchical agglomerative cluster analysis
 - Ryberg, Karen R., 2006, Cluster Analysis of Water-Quality Data for Lake Sakakawea, Audubon Lake, and McClusky Canal, Central North Dakota, 1990-2003: U.S. Geological Survey Scientific Investigations Report 2006-5202, , 38 p., <https://pubs.usgs.gov/sir/2006/5202/>.

Cluster Analysis

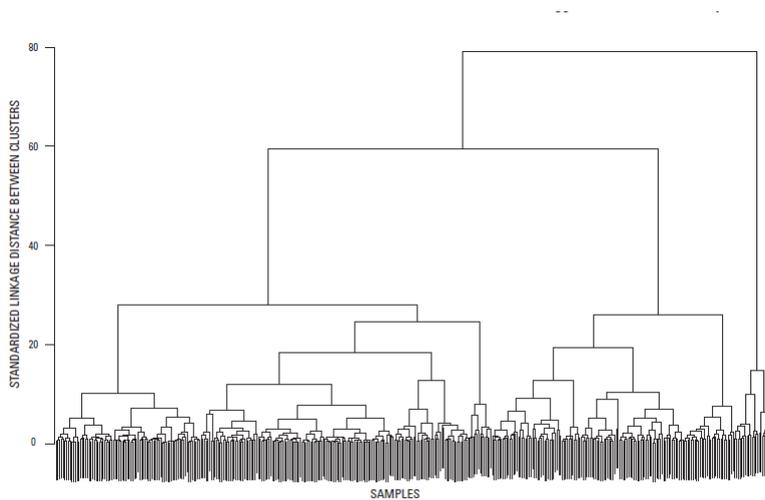


Figure 5. Dendrogram from hierarchical agglomerative cluster analysis of 409 surface-water samples collected from Lake Sakakawea, Audubon Lake, and McClusky Canal, 1990-2003. [Samples arranged so that branches of dendrogram do not cross.]

- Specific conductance
- pH
- Alkalinity
- Calcium
- Magnesium
- Sodium
- Potassium
- Sulfate
- Chloride
- Ammonia

Six Groups

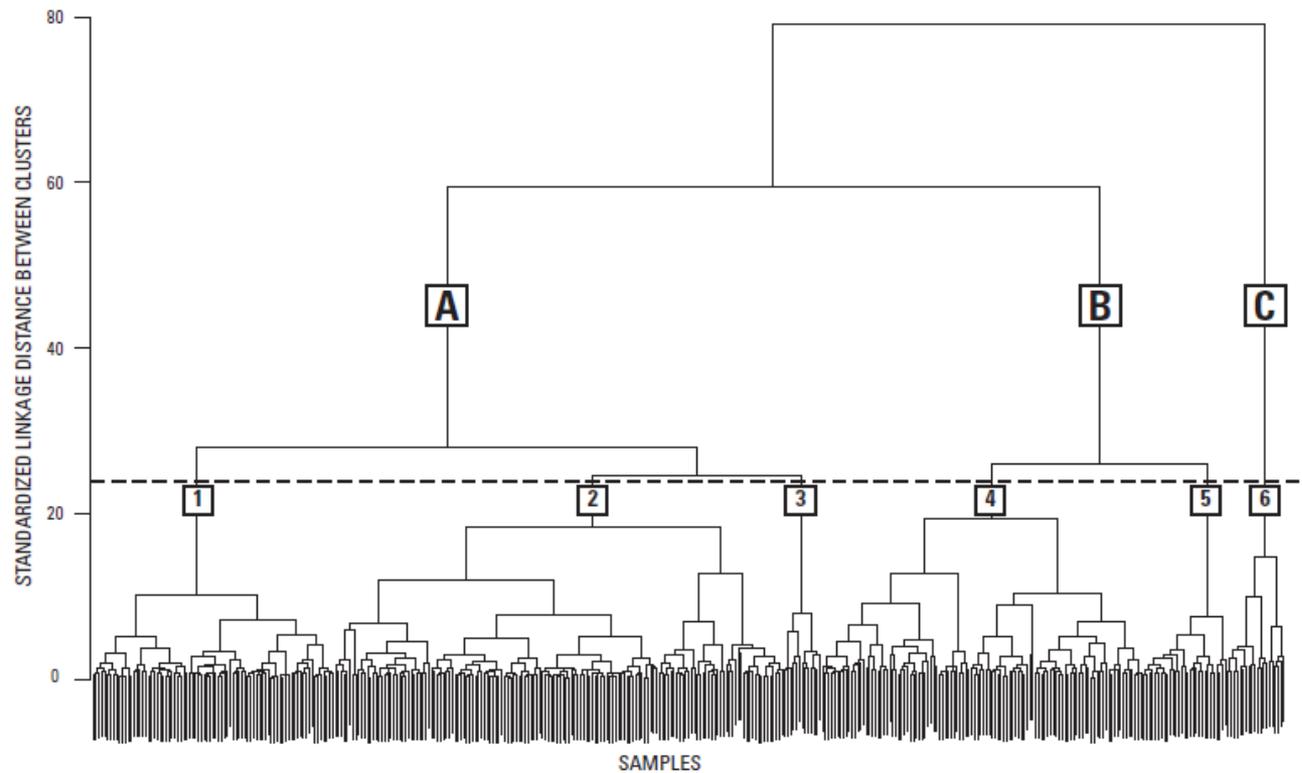


Figure 6. Dendrogram showing groups (A, B, and C) and subgroups (1-6) of surface-water samples examined in this study. The dashed horizontal line identifies the six subgroups and those clusters whose main branches extend below the line. [Samples arranged so that branches of dendrogram do not cross.]

Cluster Analysis Lead to Questions that Results in More Information about System

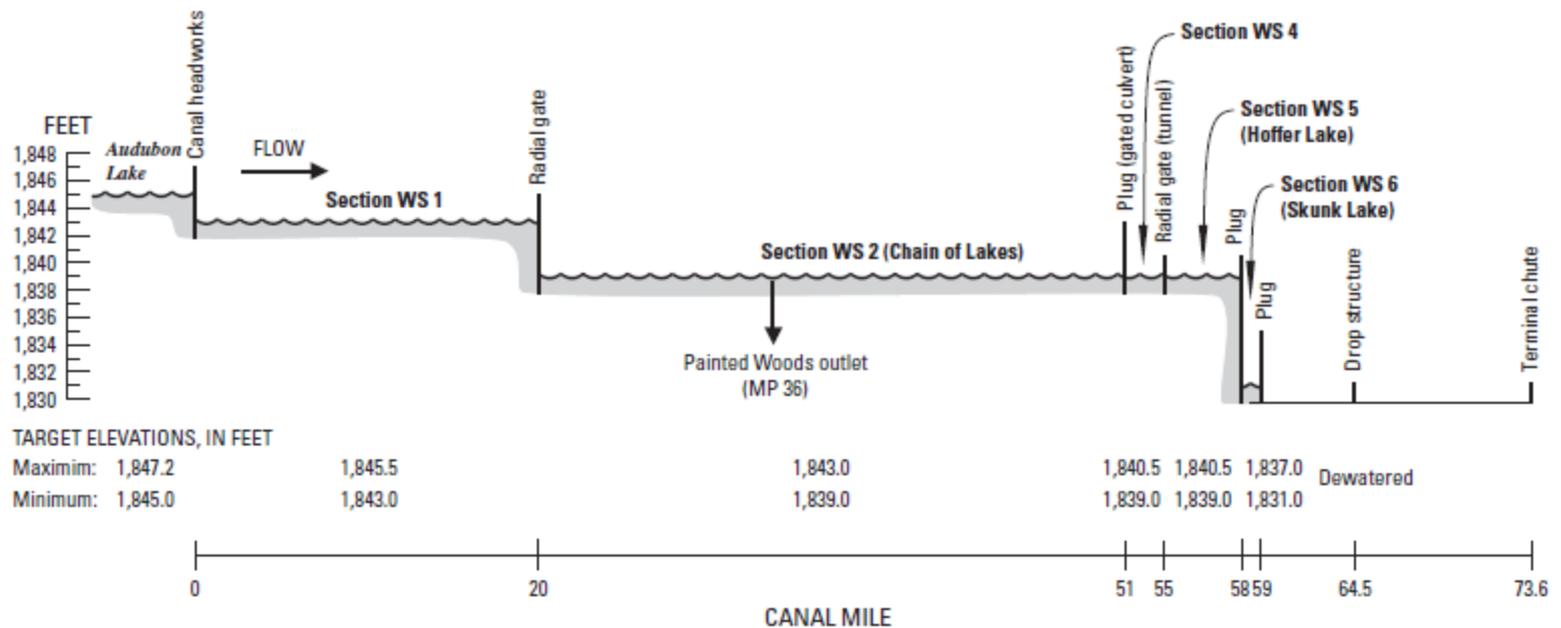


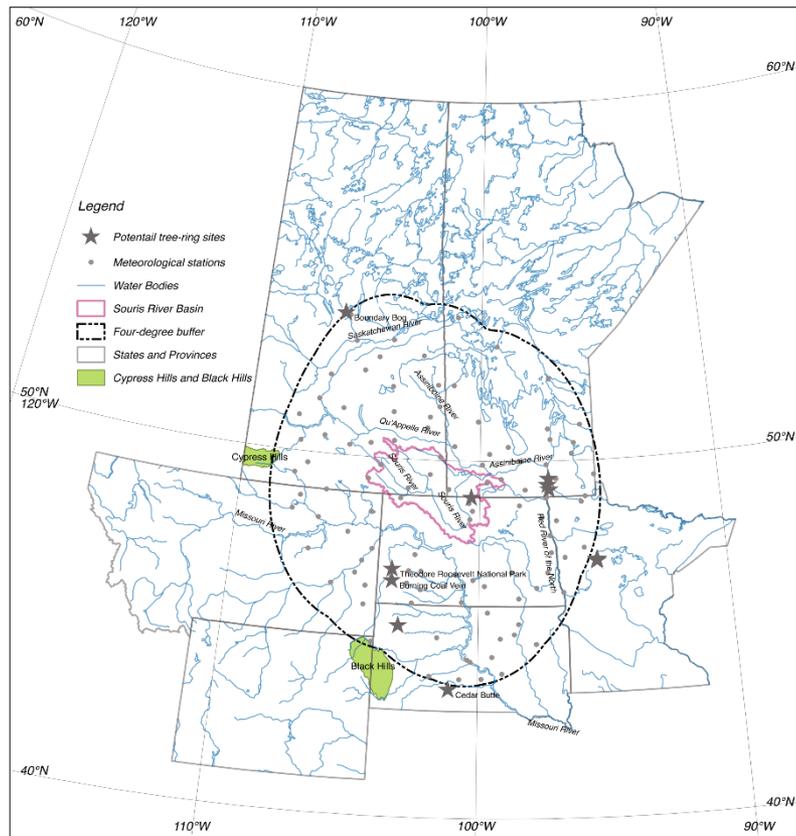
Figure 4. Schematic diagram of McClusky Canal. Section WS 2 is a combination of former WS 2 and former WS 3 (M. Marohl, Bureau of Reclamation, oral commun., 2006).



Statistically Significant Differences Among Groups

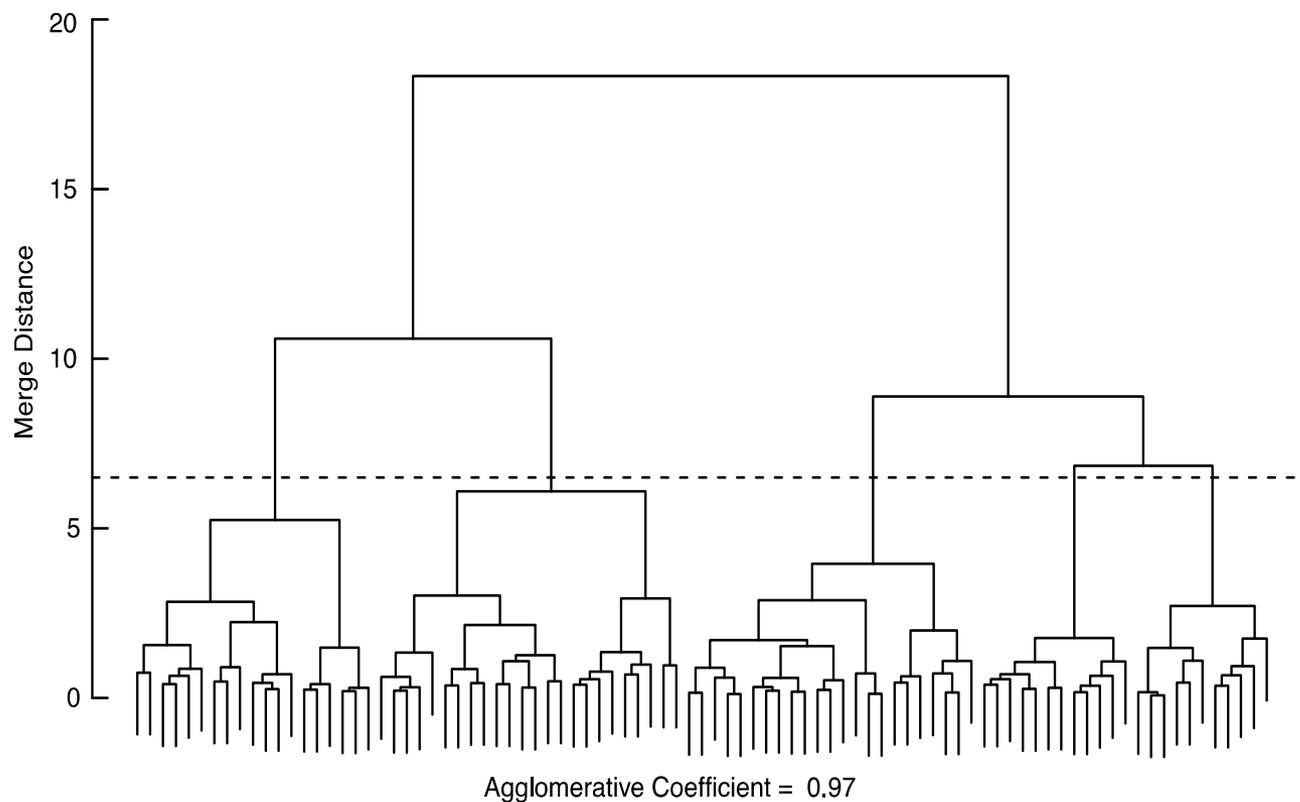
- Kruskal-Wallis rank sum test is a nonparametric test
- Test for the situation where analysis of variance (ANOVA) normality assumptions may not apply.
- Null hypothesis is the location of all the groups is the same (same median).
- Alternative hypothesis is that at least one group is different (test does not indicate which one)
- Differences can be explored graphically

Cluster Analysis



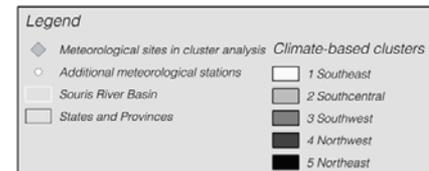
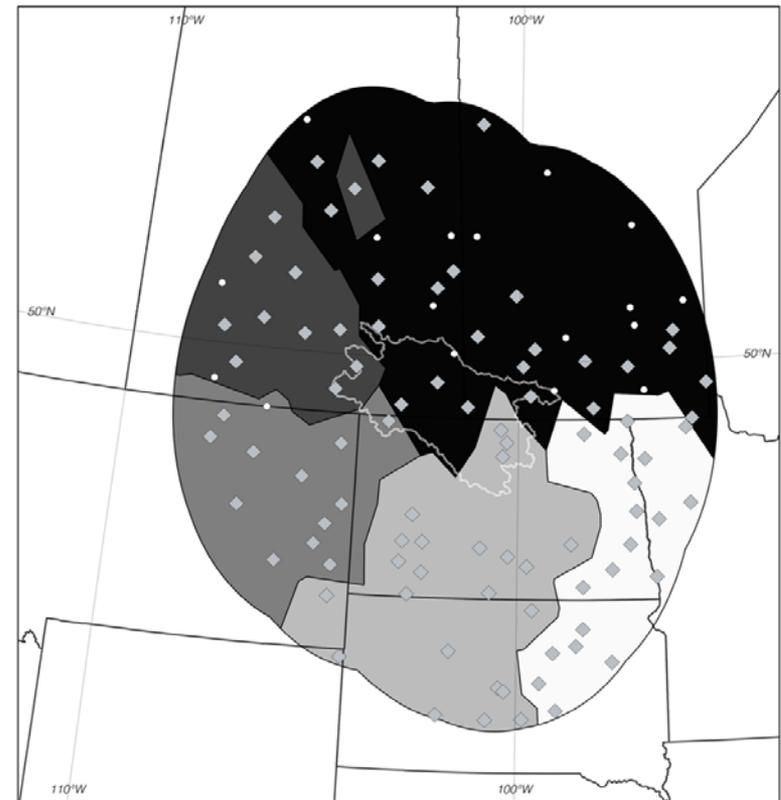
- Ryberg, K.R., Vecchia, A.V., Akyüz, F.A., and Lin, W., 2016, Tree-ring-based estimates of long-term seasonal precipitation in the Souris River Region of Saskatchewan, North Dakota and Manitoba: Canadian Water Resources Journal / Revue canadienne des ressources hydriques, 17 p., <http://dx.doi.org/10.1080/07011784.2016.1164627>
- Ryberg, K.R., 2015, The impact of climate variability on streamflow and water quality in the North Central United States: Fargo, North Dakota State University, Ph.D. dissertation, 277 p.

Hierarchical Agglomerative Clustering of Mean Seasonal Precipitation

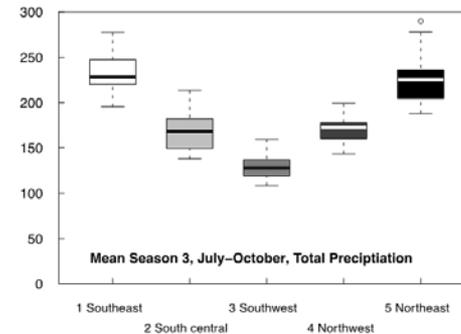
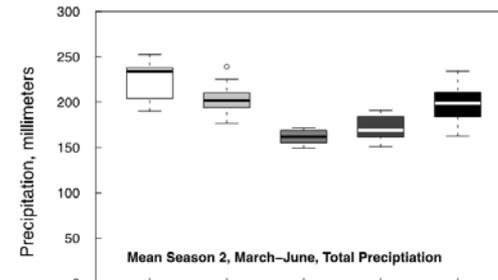
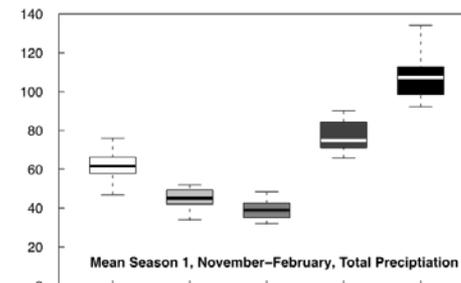
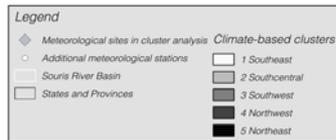


Five Regional Groups

- Application of cluster analysis and Voronoi polygons in GIS.



Graphical Examination of Differences



Principle Components Analysis

- PCA is used to explain the variance-covariance structure of a set of variables through linear combinations.
- It is often used as a dimensionality-reduction technique.
- For example, one may have many potential explanatory variables for multiple regression, but these variables are correlated and will cause multicollinearity (increases the variance of coefficient estimates, can cause instability of these estimates, incorrect signs), one could reduce the number of variables by examining how they are related with PCA.



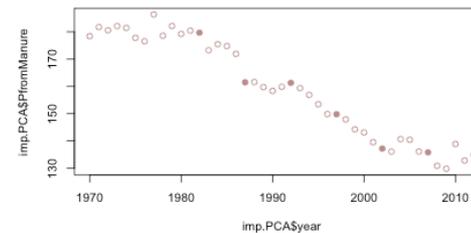
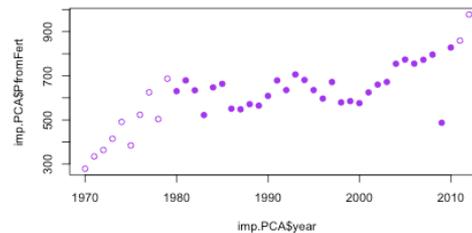
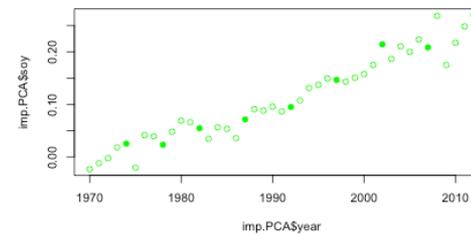
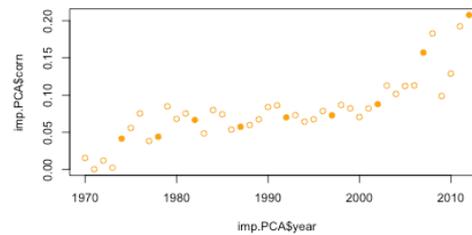
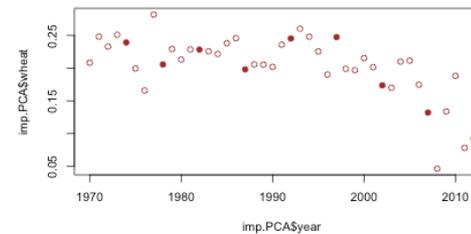
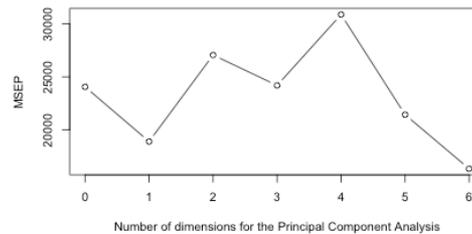
Principle Components Analysis

- See case studies in:

Sergeant, C.J., Starkey, E.N., Bartz, K.K., Wilson, M.H., and Mueter, F.J., 2016, A practitioner's guide for exploring water quality patterns using principal components analysis and Procrustes: Environmental Monitoring and Assessment, v. 188, no. 249, 15 p., doi: 10.1007/s10661-016-5253-z.

Multivariate Imputation of Missing Values

Basin 4 includes year



Resources

- The R Graph Gallery – Inspiration and Help Concerning R Graphics
 - <https://www.r-graph-gallery.com/>
- Statistical Methods in Water Resources
 - <https://pubs.usgs.gov/twri/twri4a3/>
 - New version coming in 2018
- Güler, C., Thyne, G.D., McCray, J.E., and Turner, K.A., 2002, Evaluation of graphical and multivariate statistical methods for classification of water chemistry data: Hydrogeology Journal, v. 10, p. 455-474.

Contact Information

Karen Ryberg

- kryberg@usgs.gov
- <https://www.usgs.gov/staff-profiles/karen-r-ryberg>