Exploratory Data Analysis (EDA)

ESTELLE CAMIZULI
Université Savoie Mont Blanc, France

EMMANUEL JOHN CARRANZA
University of KwaZulu-Natal, South Africa

The most common approach in statistics is to test if a chosen hypothesis is true at a given probability level (see HYPOTHESIS TESTING). This is called confirmatory data analysis (Hoaglin 2006). When determining how to properly analyze any dataset, the first consideration must be the characteristics of the data themselves (Helsel and Hirsch 2002). Exploratory data analysis (EDA) minimizes prior assumptions and guides the choice of appropriate models for further examinations (Carranza 2009; Velleman and Hoaglin 1981). This statistical approach was developed in 1977 by John Tukey, who explained that its philosophy is to look at the data to see what they seem to be saying (Tukey 1977). In simple words, one could say that EDA is used for visualization and extraction of more substantive but less obvious information from the data. Indeed, EDA uses a wide variety of techniques (descriptive statistics (see DESCRIPTIVE STATISTICS) and graphical tools) for more effective exploration of a dataset. It is robust for the study of patterns, allowing us to visualize the structure of the data and to identify outliers (errors, peculiarities, or anomalies). These should be examined carefully in order to understand the dominant behavior and the unusual behavior in the data (Hoaglin 2006). EDA leads to a better understanding about what kind of advanced statistical methods should be applied to the data (Reimann et al. 2008). EDA also questions whether the scale in which the data are originally expressed is satisfactory. If not, a transformation (e.g., logarithmic) into another scale would benefit further analysis.

To summarize a dataset, Tukey (1977) recommends using descriptive statistics with the five-number summaries based on sorting and counting. These numbers describe the data with values at selected depths: the central value is the median (Q2), the dispersion is represented by the minimum (Min) and the maximum (Max), and finally, the first quartile (Q1) and the third quartile (Q3) complete the description. The interquartile range (IQR = Q3 – Q1) measures the variation about the center of a distribution and is the most commonly used resistant measure of spread (Helsel and Hirsch 2002; VanPool and Leonard 2011). As the median is determined by the relative rank of the data, it is only minimally affected by the magnitude of a single observation. In most cases, this resistance to the effect of a change in value or presence of outliers is an advantageous property.

One way to visualize the five-number summaries is to use the Tukey box plot (Figure 1). A box is drawn from Q1 (also called the lower hinge) to Q3 (also called the upper hinge), crossing with a bar at the median. Two fences are defined according to their distance to the box. The first fences (inner fences) are one step (1.5 times IQR) outside the box, whereas the second fences (outer fences) are two steps (3 times IQR) outside the box. The whiskers extend to the value at each end closest to, but still inside, the inner fences (Tukey 1977). Values from the fences to either minimum or maximum are individually plotted so that data apparently straying out, far beyond the others, can be identified.

Box plots allow graphical representation of a dataset through the quartiles and thus provide information about the variation and central tendencies of data in a condensed manner (VanPool and Leonard 2011). It is a valuable tool to detect outliers, which are extreme values that can have several origins. They can be, according to Helsel and Hirsch (2002): (1) erroneous values (e.g., mistakes in measurements of an archaeological artifact); (2) observations from a population not similar to that of most of the data (e.g., data reflecting foreign provenance of some objects in a specific area); or (3) rare events from single populations that are skewed. It is very important to detect and examine outliers in order to

The Encyclopedia of Archaeological Sciences. Edited by Sandra L. López Varela.
© 2018 John Wiley & Sons, Inc. Published 2018 by John Wiley & Sons, Inc.
DOI: 10.1002/9781119188230.saseas0271

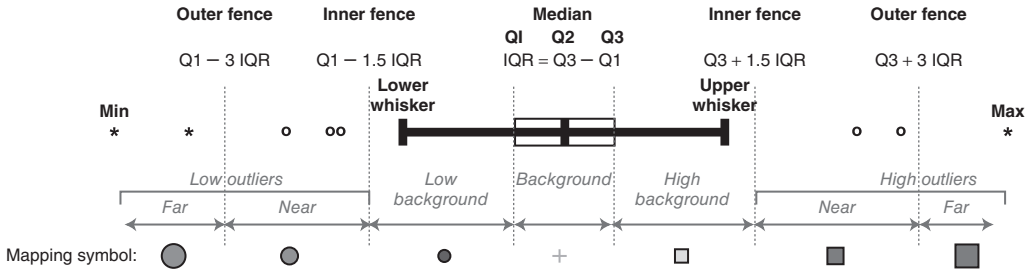

Figure 1 Tukey box plot used to describe a dataset with five-number summaries (Min, Q1, Q2, Q3, and Max). Whiskers correspond to the value at each end closest to inner fences. Seven categories based on the Tukey box plot and defined for mapping of geochemical anomalies are also specified.

determine if they should be excluded or treated with specific statistics. Box plots are excellent for the discrimination of properties within a dataset, even for comparisons between many groups of data. They are useful tools in determining whether medians, spreads, and symmetries differ among groups of data (Helsel and Hirsch 2002). Box plots, together with histograms and scatter plots are, among the EDA graphical tools, particularly used in any archaeological study.

Histograms provide a clear way of visualizing the frequency distribution of data in classes for a single variable (Figure 2, left). Bars are drawn with, on the ordinate, the number of data that fall into several categories on the abscissa. The histogram exhibits data structure, the central values, and the range per class. However, the visual display depends on the choice of the number of classes. The histogram permits at a single glance the detection of whether a data distribution is symmetrical or skewed. It is also readily apparent from a histogram if the data distribution is unimodal (with one peak depicting the presence of a single population) or multimodal (with several humps depicting the presence of several populations) (Reimann et al. 2008).

A two-dimensional scatter plot is an informative and simple graphic to study the relationships between variables (Figure 2, right). In EDA, data for two variables can be plotted to identify any unusual structures in the data (linear/curved, separate regions, variability in spread) and assist in determining which process could produce these unusual structures (Helsel and Hirsch 2002; Reimann et al. 2008).

Box plots are useful for comparing data for single variables, whereas two-dimensional scatter plots illustrate relationships between two variables. However, it is also important to study relationships when there are more than two variables (which is often the case in archaeology). Several other tools/methods of EDA can provide insight into variable relationships, such as scatter plot matrix (Figure 2, center) and principal components analysis (see PRINCIPAL COMPONENT ANALYSIS) (PCA). PCA is a method to express differently the data. The PCA-transformed data are a linear combination of the original variables, uncorrelated and arranged in order of decreasing variances (Helsel and Hirsch 2002). PCA is dependent on the scale, so the transformation of data before analysis is of primary concern. This method is used to study correlations between the variables in large datasets, and to find which variables are redundant for further analysis. PCA helps to distinguish different populations and allows the identification of outliers. It is often used before a cluster analysis.

The EDA approach is widely used in various scientific domains such as environmental sciences, earth sciences, and economics, as well as in all the fields of archaeology. The work of archaeologists is indeed to analyze archaeological records (see ARCHAEOLOGICAL RECORD), behind which is the idea of describing, counting, and measuring these records, to categorize them in order to interpret function, provenance, or to a greater extent human behavior and culture (see ARCHAEOLOGICAL CLASSIFICATION). In other words, one of the archaeologist's objectives is to study the typology of artifacts through qualitative (e.g., color estimation) and quantitative (e.g., length and width measurements) variables. These typological studies can lead to the

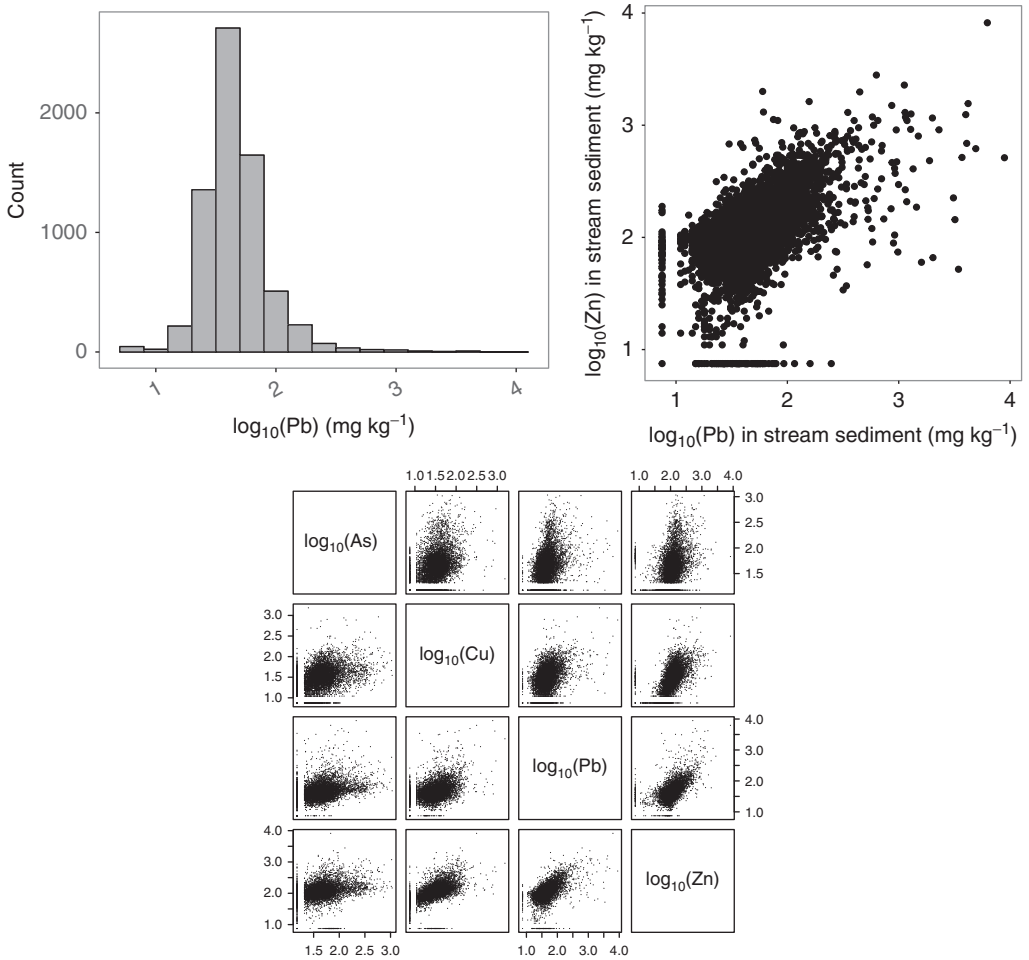

Figure 2 Three examples of EDA graphical tools: histogram (left), two-dimensional scatter plot (right), and scatter plot matrix (center).
Source: Geochemical data from the French geological survey (http://sigminesfrance.brgm.fr/). Graphics were created using R software (R 3.3.0, https://www.R-project.org/).

collection of large datasets. EDA is also useful when data are not sufficient for confirmatory data analysis, and this is important because sometimes archaeologists have to deal with limited/subset data (i.e., the study materials are only a portion of the past) (VanPool and Leonard 2011).

EDA applications in archaeological fields are numerous. EDA of the spatial distribution of chemical elements in archaeological soils helps to understand the function of specific areas. In anthropology and archaeozoology, the analysis of morphological traits may help to study population movement. To study ceramic or lithic provenance, for example, with trace element and mineralogical analyses, archaeologists have to define potential subgroups. This can be done by combining several EDA tools such as bivariate plots, cluster analysis (see CLUSTER ANALYSIS), or compositional profile plots. These tools assist archaeologists in the first step to identify artifact composition groups and to partition a dataset into groups of samples that apparently have distinctive compositions.

An application of EDA is detailed here for the interpretation of trace metal anomalies as a tool for field prospection in mining archaeology. This case study illustrates the potential of EDA to assist explaining how geochemical anomalies in streambed sediments could help in tracing ancient mining sites. The study area is the French northern Alps where several phases of mining occurred from the Bronze Age to the twentieth century. The mining sites are spread all over the territory and various substances were exploited including iron and lead (Figure 3, left). The exploitation of mineral deposits may considerably extend mineral dispersal, thus causing persistent

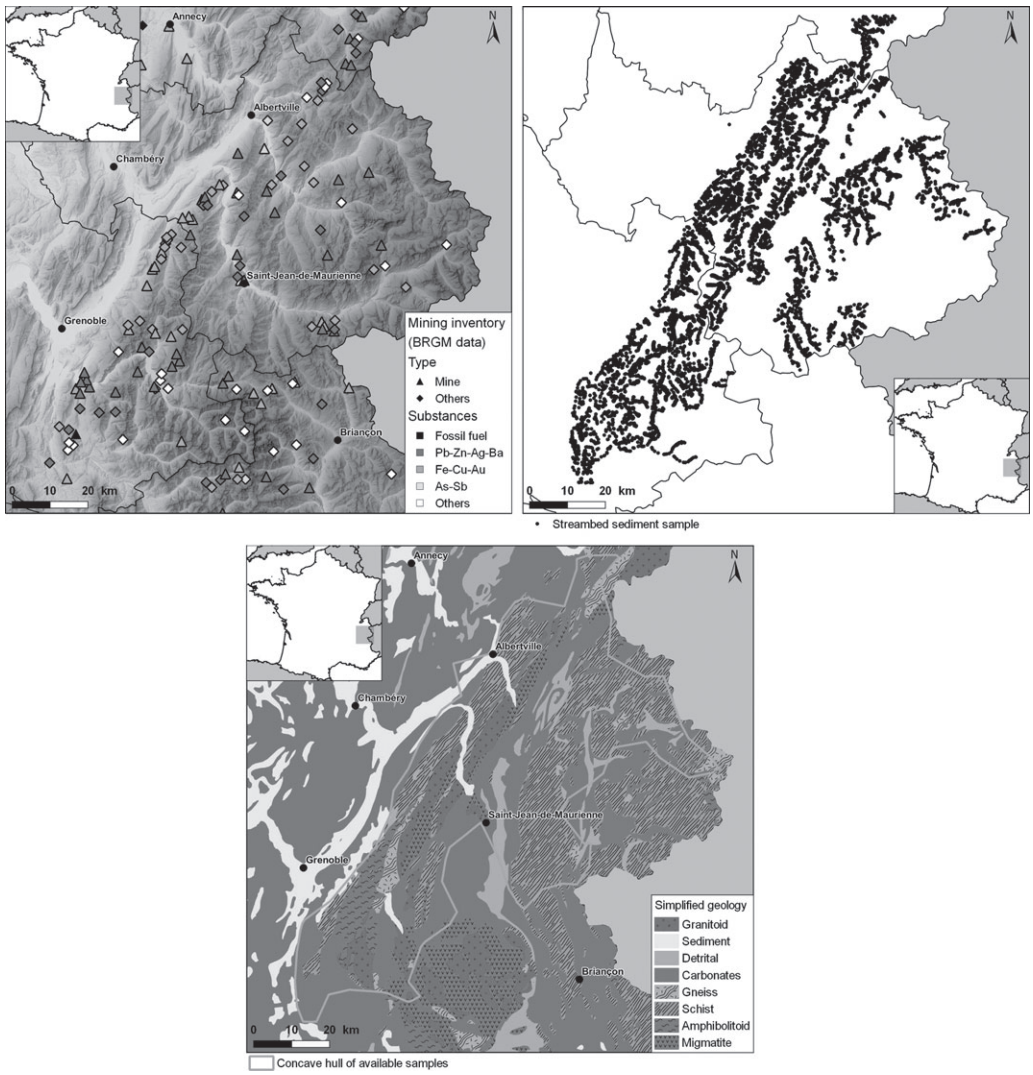

Figure 3 Inventory of mining sites in the French northern Alps (left), location of streambed sediment samples in the same area (right), and simplified map of lithological units (center).
Source: Data (geology, mines, and sediments) are from the French geological survey. Databases for the mines and streambed sediment samples are available at http://sigminesfrance.brgm.fr/. The maps were created using QGIS software (QGIS Essen 2.14.6, http://www.qgis.org), adapted by E. Camizuli from © IGN-2017.

local trace metal anomalies. The underlying idea of this case study is that streambed sediments represent composite erosion products of terrains outcropping in the catchment area. There are thus ideal proxies for studying the spatial repartition of ancient mining sites according to geochemical anomalies.

In the 1980s, the French geological survey (BRGM) performed a national inventory of mineral substances (http://sigminesfrance.brgm.fr/), leading to a huge amount of geochemical data. In the French northern Alps, 6,892 streambed sediment samples were collected (Figure 3, right), sieved at 125 µm, dried, and analyzed by direct current plasma for elemental compositions. For this case study, we focus on lead (Pb) concentrations because this element is known as a persistent pollutant, causing harmful effect (Kabata-Pendias and Mukherjee 2007). Since the first evidence of mining and metallurgical activities, large amounts of metals including Pb have been emitted into the environment. Lead is known for its low mobility in the environment and can be trapped in soils and sediments for centuries; this is thus an excellent choice to study the mining anomalies. Moreover, Pb is a chemical element commonly studied in archaeology, particularly in lake sediments or metal artifacts. The ratio of Pb isotopes is indeed used for dating, studying provenance, and identifying pollution sources.

Geological information is of prime importance as well for determining correlation of streambed sediment samples with underlying local lithology. Simplified lithological units in the study area were extracted from the 1:1,000,000 scale map of the BRGM (Figure 3, center).

The point is to determine threshold trace metal concentrations by EDA procedures so as to distinguish anomalies from the local geochemical background. Such anomalies could be related to certain lithology but also to mining activities. A transformation procedure was applied to the data before computing thresholds, to make comparable the streambed sediment samples coming from different lithological units. In short, the data were centered (subtraction) with respect to the median (*med*) and then scaled (division) with respect to

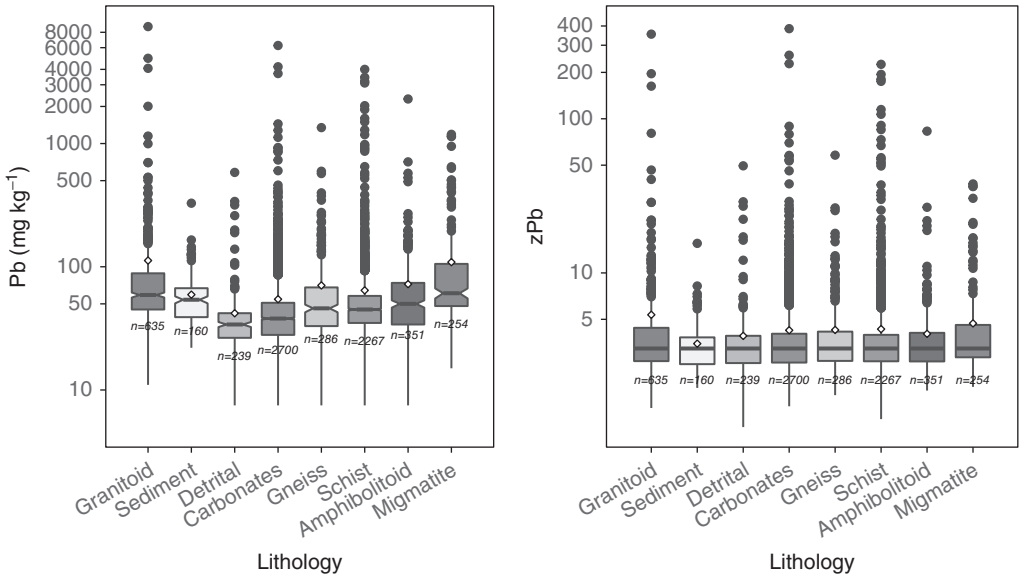

Figure 4 Pb concentrations in streambed sediments in the French northern Alps according to simplified lithological units.
Source: Original data (left) and standardized data (right) for EDA mapping (lithological data derived and modified from French geological survey 1:1,000,000 scale map). Graphics were created using R software (R 3.3.0, https://www.R-project.org/).

the median absolute deviation (MAD) for each lithology (*j*), thus:

$$Z_{Pb,j} = \left(X_{Pb,j} - med_{Pb,j}\right) \; / \; MAD_{Pb,j}$$

MAD is another resistant estimator of spread and is somewhat equivalent to the standard deviation (Helsel and Hirsch 2002). However, with the kind of transformation using the equation above, the transformed values can be negative and thus cannot be log-transformed. Another transformation was then applied to make all values greater than or equal to 1:

$$z_{Pb,j} = 1 \; + \; Z_{Pb,j} - min\left(Z_{Pb,j}\right)$$

Figure 4 illustrates the use of the above equations to standardize Pb concentrations (zPb) in streambed sediments according to the lithological units.

Once the data are comparable, six Tukey thresholds were computed defining seven categories for EDA mapping: far and near low outliers, low background, background, high background, near and far high outliers (Table 1, Figure 1). The thresholds were calculated using the map.eda7 function of the rgr package, supplementing the R software (R Development Core Team 2016).

Near and far high outliers are the two categories that are particularly interesting for highlighting ancient mining site areas. Computation shows that these categories represent 5.04 percent of the dataset (Table 1). It is easy to link each sample to the corresponding category, and, as each sample possesses coordinates, to

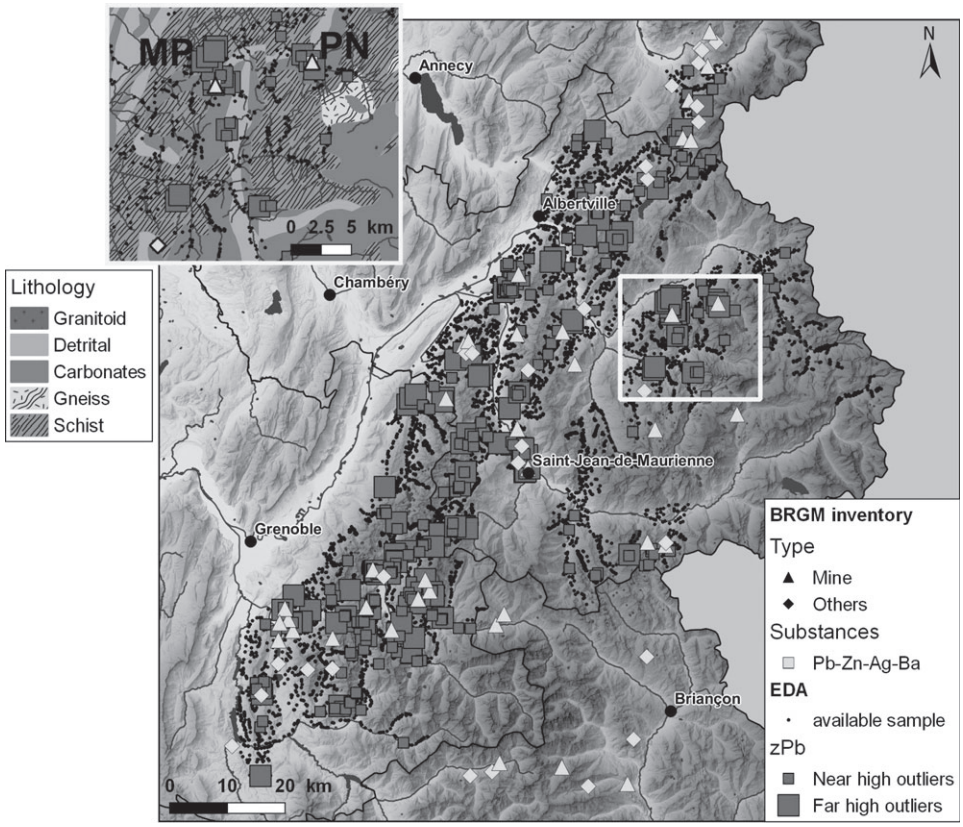

Figure 5 Mapping of high EDA anomalies in standardized Pb concentration (zPb), with data from streambed sediments in the French northern Alps. Also shown are Pb mining sites inventoried by the French geological survey. MP, Macôt-la-Plagne; PN, Peisey-Nancroix.
Source: The maps were created using QGIS software (QGIS Essen 2.14.6, http://www.qgis.org), adapted by E. Camizuli from ©IGN-2017.

Table 1 Ranges of the seven categories for EDA mapping, calculated from standardized Pb concentrations in streambed sediment samples, and their respective percentage in the whole dataset.

EDA category	*Range (zPb)*	*Percentage*
Far low outliers	≤ 0.75	0.00
Near low outliers	(0.75–1.41)	0.73
Low background	(1.41–2.65)	24.23
Background	(2.65–4.03)	49.58
High background	(4.03–7.5)	20.43
Near high outliers	(7.56–14.2)	3.34
Far high outliers	> 14.2	1.70

map them with the mining inventory (Figure 5). When focusing on an area with known mining activities, such as Macôt-la-Plagne (eighteenth to twentieth centuries) and Peisey-Nancroix (eighteenth to nineteenth centuries), it is easy to visualize their impacts. Areas with anomalies and unknown mining sites would be good candidates for prospecting if no other explanation for their presence can be found.

EDA is widely used by archaeologists to summarize and graphically display their data. These procedures and confirmatory data analysis complement each other. EDA results may stimulate further questions and lead to further analysis. If possible, data should be investigated with both approaches to avoid misinterpretation (Reimann et al. 2008).

SEE ALSO: Nonparametric Statistics; Statistics and GIS; Statistics in Archaeology

REFERENCES

Carranza, Emmanuel John M. 2009. "Geochemical Anomaly and Mineral Prospectivity Mapping in GIS." In *Handbook of Exploration and Environmental Geochemistry*, Vol. 11, edited by M. Hale. Amsterdam: Elsevier.

Helsel, Dennis R., and R. M. Hirsch. 2002. *Statistical Methods in Water Resources*. In *Techniques of Water-Resources Investigations. Book 4: Hydrologic Analysis and Interpretation*, Chapter A3. Reston, VA: US Geological Survey.

Hoaglin, David C. 2006. "Exploratory Data Analysis." In *Encyclopedia of Statistics*, 2nd ed., edited by Samuel Kotz, N. Balakrishnan, Campbell B. Read, Brani Vidakovic, and Norman L. Johnson, 2151–55. Hoboken, NJ: John Wiley & Sons.

Kabata-Pendias, Alina, and Arun B. Mukherjee. 2007. *Trace Elements from Soil to Human*. Heidelberg: Springer.

R Development Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing, accessed February 8, 2018. http://www.R-project.org.

Reimann, Clemens, Peter Filzmoser, Robert G. Garrett, and Rudolf Dutter. 2008. *Statistical Data Analysis Explained: Applied Environmental Statistics with R*. Chichester, UK: John Wiley & Sons.

Tukey, John Wilder. 1977. *Exploratory Data Analysis*. Addison-Wesley Series in Behavioral Science: Quantitative Methods. Reading, MA: Addison-Wesley.

VanPool, Todd L., and Robert D. Leonard. 2011. *Quantitative Analysis in Archaeology*. Chichester, UK: Wiley-Blackwell.

Velleman, Paul F., and David C. Hoaglin. 1981. *Applications, Basics, and Computing of Exploratory Data Analysis*. Boston, MA: Duxbury Press, accessed February 8, 2018. http://ecommons.cornell.edu/handle/1813/78.

FURTHER READING

Monna, Fabrice, Estelle Camizuli, Rachid Nedjai, Florence Cattin, Christophe Petit, Jean-Paul Guillaumet, Isabelle Jouffroy-Bapicot, Benjamin Bohard, Carmela Chateau, and Paul Alibert. 2014. "Tracking Archaeological and Historical Mines Using Mineral Prospectivity Mapping." *Journal of Archaeological Science* 49 (September): 57–69. DOI:10.1016/j.jas.2014.04.022.