DATA ACCURACY AND VALIDATION

N. B. Harmancioglu
Department of Civil Engineering, Faculty of Engineering, Dokuz Eylul University, İzmir, Turkey

V. P. Singh
Department of Civil Engineering, Louisiana State University, Baton Rouge, LA, USA

Keywords: Data, information, data management system, data accuracy, data reliability, data validation, quality assurance, quality control, calibration, noise, random error, systematic error, homogeneity, consistency, stationarity.

Contents

Bibliography
Biographical Sketches

Summary

Recently, recognition of the gap between information provided by available data and that required for environmental management has brought focus to current monitoring systems, databases, data validation and data use. Accordingly, major efforts have been initiated at regional and international levels to improve the status of existing information systems. The purpose of these efforts is to ensure that the data made available to users are accurate and reliable.

Data are transferred into information via a data management system that involves a number of steps comprising data acquisition, processing and the eventual data analyses for preparation of operational and design data. Each of these steps contributes to the retrieval of the required information and has an impact on the quality of data collected and processed. Thus, all of these steps must be efficient to maximize data utility and reliability, meaning that quality controls (QC) should be realized at each step. In particular, it is necessary that collected data are validated before they are disseminated to users. The users themselves can apply a number of checks to test whether the data are representative of the environment before they use them as a basis for their operational and design decisions.

To comply with the requirements imposed on current information systems, this article focuses on problems associated with data accuracy and reliability and discusses possible means of validating available data. Shortcomings of current environmental data are described within the framework of a data management system. Data validation is considered within the scope of both the data processing activities and the user's final check for representativeness of data.

1. Introduction

Agenda 21 of UNCED (United Nations Conference on Environment and Development, Rio de Janeiro, 1992) has officially stated the new outlook towards environmental management, namely that the environment should be managed by an integrated approach in respect of sustainability. It was further emphasized in Agenda 21 that effective management relies essentially on reliable, accurate, and adequate information on how the environment behaves under natural and man-made impacts. Chapter 40 of Agenda 21 on Information for Decision-Making focuses particularly on **informed decision-making** for environmental management and stresses the **need to ensure that decisions are based increasingly on sound information**.

On the other hand, Agenda 21 and several other international documents and reports have also recognized that current systems of information production, that is, data management systems, do not fulfill the requirements of environmental management and decision-making. In view of the rapidly growing environmental problems, it is often found that our data management systems experience a declining trend at a time when informational support is needed the most. There is a significant gap between information needs on environment and information produced by current systems of data

collection and management. The presence of this gap contradicts the nature of the Information Age we live in.

Recognition of the gap between information provided by available data and that required for environmental management has brought focus to current monitoring systems, databases, data validation and data use. Accordingly, major efforts have been initiated at regional and international levels to improve the status of existing information systems. The purpose of these efforts is to ensure that the data made available to users are accurate and reliable.

Data are transferred into information via a data management system that involves a number of steps comprising data acquisition, processing and the eventual data analyses for preparation of operational and design data. Each of these steps contributes to the retrieval of the required information and has an impact on the quality of data collected and processed. Thus, all of these steps must be efficient to maximize data utility and reliability, meaning that quality controls (QC) should be realized at each step. In particular, it is necessary that collected data are validated before they are disseminated to users. The users themselves can apply a number of checks to test whether the data are representative of the environment before they use them as a basis for their operational and design decisions.

Despite the above requirements, each step of a data management system is subject to numerous uncertainties and difficulties so that shortcomings are often encountered in available data. These shortcomings relate to the reliability, accuracy, completeness (missing values), homogeneity, length of record and spatial extent of data. There are often no measurements of sampling error indicated along with available data. In particular, data validation is often poorly achieved. The result is that the eventual information produced is of poor quality, imprecise and unreliable. Decisions based on such information are prone to significant errors such that management of the environment cannot be realized in an efficient and cost-effective manner.

To comply with the requirements imposed on current information systems, this article focuses on problems associated with data accuracy and reliability and discusses possible means of validating available data. Shortcomings of current environmental data are described within the framework of a data management system. Data validation is considered within the scope of both the data processing activities and the user's final check for representativeness of data.

2. Data Requirements for Environmental Management

2.1 Types of Data Needed

Substantial amounts of data already exist on various processes occurring in the natural environment. However, the mode of adoption of integrated approaches for sustainable development of the environment has certainly changed information expectations and, hence, the types and the amounts of data needed. Now, more and different types of data have to be collected to describe the status and trends of the ecosystem, natural resources, pollution, and socioeconomic variables. As current environmental problems

extend to freshwater (both surface and groundwater), land resources, coastal zones, urban air, desertification, soil degradation, biodiversity, and other habitats, data are required on all these media so that such problems can be assessed and managed.

Considering freshwaters, conventional water resources information systems comprise hydrological and meteorological data on such processes as precipitation (rainfall, snow), river levels and flows, lake and reservoir levels, groundwater levels, sediment concentrations and loads in rivers, evapotranspiration, and water quality (physical, chemical, and bacteriological variables) of surface and groundwater. On the other hand, freshwaters are now considered a part of the environmental continuum comprising air, soil and water components that are interactive in complex ways. Thus, there is now a need to collect data on the wider environment to include watershed characteristics such as vegetation patterns, soil moisture, topography, climate, and aquifer characteristics. Environmental data should include a wide variety of variables to provide information on diffuse sources of pollutants, accidental spills, irrigation return flows, eutrophication of lakes, status of estuarine and coastal ecosystems. Such data essentially reflect human impact on the natural environment. In a similar vein, data are also needed to describe water use by man, i.e., the volumes of water required for domestic, industrial and agricultural use, and characteristics of rivers related to catchment area uses such as recreation, navigation and fishery habitats.

It is clear from the above that the types of data required to produce information on the environment are highly varied. In addition, these data should reflect the true nature of the environment. Environmental processes are, by nature, heterogeneous, dynamic, nonlinear and anisotropic. They are marked by spatial variability as well as temporal variability. Accordingly, collected data should reflect these characteristics of the environment along with the spatial and temporal variability of environmental processes to be representative of nature.

2.2 Statistical Characteristics of Environmental Data

Information conveyed by data may be presented in the form of plots or tabulated data. On the other hand, statistics are used to express the data in a summary form to describe either the status or the changing conditions (trends) of an environmental process.

The statistical information that available data provide include:

(a) mean values (annual, monthly, seasonal or daily);
(b) extreme values (maxima and minima) and selected percentiles;
(c) measures of variability (standard deviation, variance or coefficient of variance);
(d) continuous records such as flow hydrographs;
(e) trends (spatial and/or temporal);
(f) periodic fluctuations;
(g) auto-correlation.

Each of above properties needs different data analysis techniques so that it can be reliably described. Such techniques are further classified according to their suitability to

the nature of available data. Some methodologies require regularly collected data; whereas some can better adapt to sporadically (irregularly) observed data series.

Environmental processes may be analyzed as univariate series in the form of either time series (as a function of time) or as line series (as a function of distance) when one of the dimensions, time or space, is kept constant. However, information is often needed on both the temporal and spatial distribution of such processes so that one has to consult to multivariate analysis techniques for a full understanding of how they evolve over time and space. In this sense, rainfall and runoff data are probably the least problematic as they are regularly observed within a systematically operated network. Water quality data, on the other hand, pose significant difficulties in multivariate analyses due to the monitoring practice applied.

2.3 The Use of Statistics in Environmental Assessment

2.3.1 Identification of Driving Forces

Driving forces provide the input to the watershed. The input can be natural, or man-made such as acid precipitation, both point source and non-point source, such as waste discharge from an industry, city sewage water, agricultural pollution due to chemical fertilization, etc. The data expressing the driving forces must be checked for quality, trend, completeness, homogeneity or consistency. Frequently, there are gaps in the data, and they must be filled in. Mass curves are used for checking the homogeneity or consistency of data. Normal ratio method, inverse distance squared method, and correlation method are among the methods for filling in missing values. Entropy method is also used for this purpose. The data must also be checked for their errors, representativeness, and sampling strategy. All data are not collected at the same temporal frequency. Some are collected more frequently than others. The question then arises: How to transform them to the same base frequency without undue loss of information. Statistics help accomplish this objective. Statistical methods for trend detection are employed if data occur with persistence.

2.3.2 Watershed System

When dealing with a watershed, soil, vegetation, land use, morphology and geology must be known or specified. These characteristics influence pollutant transport and storage within the watershed. A complicated specification of these characteristics is their spatial variability. Statistics aid in characterizing this variability. Furthermore, statistics help classify basins based on similarity and homogeneity measures. These latter measures are useful when transposing results from one watershed to another. Here, the correlation methods and kriging are helpful.

2.3.3 Observed Data

Once a monitoring network has produced some data, these data are analyzed for information extraction. Using these data, the network design is checked for optimality. In other words, if sampling frequencies in space and time are acceptable and are in accord with the design objectives, and the cost of data collection is not prohibitively

large, then the designed network is satisfactory. To that end, entropy and correlation methods are employed. Other statistical measures, such as spectral methods and information content, can also be used.

2.3.4 Selection of Methods

Statistics give criteria to check robustness of methods and then predicate the basis for selection of suitable models. The criteria most frequently employed are bias, root-mean-square error, and coefficient of efficiency. The first two are added to define a robustness criterion.

2.3.5 Modeling Techniques

Environmental data are essential for environmental management as well as for model building, calibration, verification and real-time application. Data requirements of different models are, however, different since these models are intended for different purposes. On the other hand, depending on the availability of the type and quality of data, different types of models are developed. Thus, data and models are interdependent.

In practice, two criteria can be distinguished by which models and their data requirements are identified: (1) spatial and temporal resolution, and (2) level of analysis. Two broad categories of temporal resolution include **continuous** and **discrete**, and those of spatial resolution include **lumped** and **distributed** (or spatially continuous) types. The level of analysis is determined by the amount and resolution of the data available (both quantity and quality) on one hand and by the purpose of the assessment and availability of resources on the other.

3. Environmental Data Management

3.1 Need for Data Management

Data availability is not a sufficient condition to produce the required information about environment. It is the utility or usefulness of data that contributes to production of information. In the past, the primary concern was to conceive what available data showed about prevailing conditions of the environment. The question nowadays is whether the available data convey the expected information.

Data collection systems have indeed become sophisticated with new methods and technologies. However, when it comes to utilizing collected data, no matter how numerous they may be, one often finds that available samples fail to meet specific data requirements foreseen for the solution of a certain problem. In this case, the data lack utility and cannot be transferred into the required information. This is one of the reasons why data systems must be managed; that is, data management is required to produce an efficient information system where data utility is maximized.

Another aspect of the problem lies in cost considerations. Data collection and dissemination are costly procedures; they require significant investments that have to be amortized by versatile uses of data. Even in the developed countries, a data collection

system has to be realized under the constraints of limited financial sources, sampling and analysis facilities, and manpower. If the output of this system, or the data, do not fulfill information expectations, the investment made on the system cannot be amortized so that the result will inevitably be economic loss.

Cost considerations do not only relate to costs of monitoring; they are also reflected in the eventual decision-making process. If available data produce the required information, decisions are made more accurately, and the smaller the chances are of under-design and over-design. Proper decisions minimize economic losses and lead to an overall increase in the benefit/cost ratio. Thus, a data collection system has to be cost-effective and efficient to avoid economic losses both in the monitoring system itself and in the eventual design based on the information produced by this system.

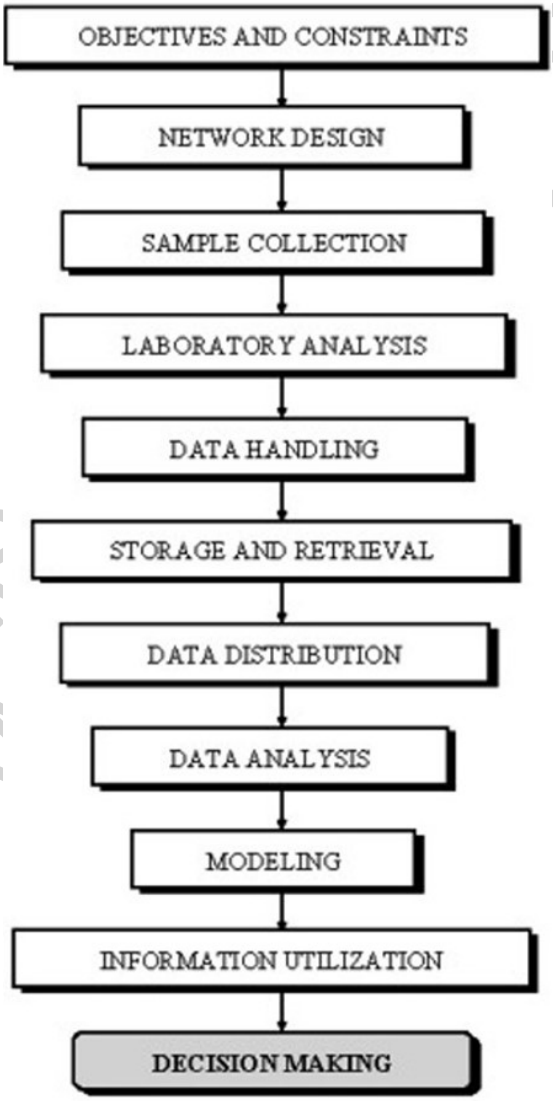

Figure 1. Basic steps in environmental data management.

The transfer of data into information involves several activities in sequence as summarized in Figure 1. Each of these activities contributes to retrieval of the required information. Thus, all of these steps must be efficient to maximize data utility and reliability. To respect the condition of cost-effectiveness, again each step has to be economically optimized. Thus, these activities have to be managed to ensure the efficiency and cost-effectiveness of the whole information system. At present, a further requirement is imposed on data management systems, namely that they should be evaluated via integrated approaches.

-
-
-

Bibliography

Clark M. J. R. and Whitfield P. H. (1994). Conflicting perspectives about detection limits and about the censoring of environmental data, *Water Resources Bulletin* **30**(6), 1063–1079. [This work discusses QA/QC procedures for validation of environmental data and describes sources of random and systematic errors in data.]

Finlayson B. L. and McMahon T. A. (1995). Understanding river hydrology. *Environmental Hydrology* **15** (ed. V. P. Singh), pp. 107–135. Dordrecht: Kluwer Academic Publishers, Water Science and Technology Library, Vol. 15. [This chapter addresses data validation and assessment of errors in runoff data records.]

Harmancioglu N. B. Fistikoglu O., Ozkul S. D., Singh V. P., and Alpaslan, M. N. (1999). *Water Quality Monitoring Network Design* 290 pp. Dordrecht: Kluwer Academic Publishers, Water Science and Technology Library, Vol. **33**. [This presents basic guidelines to be applied in water quality monitoring network assessment and redesign.]

Harmancioglu N. B., Singh V. P., and Alpaslan M. N., eds. (1998). *Environmental Data Management*, 298 pp. Dordrecht: Kluwer Academic Publishers, Water Science and Technology Library, Vol. **27**. [This book focuses on all activities involved in a data management system for transfer of data into information.]

Harmancioglu N. B., Alpaslan M. N., Ozkul S. D., and Singh V. P., eds. (1997). *Integrated Approach to Environmental Data Management Systems*, 546 pp. Dordrecht: Kluwer Academic Publishers, NATO ASI Series, 2. Environment, Vol. **31**. [This presents the outcomes of a NATO Advanced Research Workshop, which address all aspects of a data management system for purposes of integration in environmental data management on an international and multidisciplinary basis.]

Singh V. P., ed. (1995). *Environmental Hydrology*, 479 pp. Dordrecht: Kluwer Academic Publishers, Water Science and Technology Library, Vol. **15**. [This book discusses a unified approach to the role of hydrology in environmental planning and management, focusing on data requirements for environmental management, data analysis and modeling.]

Timmerman J. G., Gardner M. J., and Ravenscraft J. E. (1996). *Quality Assurance*, 119 pp. Lelystad: UN/ECE Task Force on Monitoring and Assessment, Vol. **4**, RIZA Report No. 95-067. [This report presents an exhaustive survey of sources of noise in data and their treatment with a focus on QA/QC aspects.]

UN (1992). *Agenda 21: Programme of Action for Sustainable Development, Chapter 40 on Information for Decision-Making*. New York: United Nations. [This focuses on data requirements for informed decision-making in environmental management in the 21st Century.]

WMO/UNESCO (1991). *Report on Water Resources Assessment*, 65 pp. Oxford: Words and Publications. [This report delineates data needs for decision-making in environmental management.]

Yevjevich, V. (1972). *Probability and Statistics in Hydrology*, 302 pp. Fort Collins: Water Resources Publications. [This book provides information on data characteristics and statistical methods for detection and correction of errors in data.]

Biographical Sketches

Nilgun B. Harmancioglu is a professor of hydrology and water resources, teaching and carrying out research activities at Dokuz Eylul University (DEU) Faculty of Engineering since 1976. She received her B.Sc. in civil engineering in 1973, M.Sc. in water resources engineering in 1976, and her Ph.D. in hydrology and water resources in 1981. She was promoted to associate professorship in 1986 and to full professorship in 1992 at DEU Faculty of Engineering. Currently, she is the head of the Hydraulics, Hydrology and Water Resources Division of the same faculty. Dr Harmancioglu spent a year between 1980-81 in Paris, Ecole National Superieure des Mines de Paris, Centre d'Informatique Geologique, Laboratoire d'Hydrogeologie Mathematique for post-doctoral studies on optimum design of hydrometric data collection networks. Between 1984-86, she worked as a research associate in Washington, D.C., the George Washington University, School of Engineering and Applied Science, International Water Resources Institute. There she carried out a research project on monitoring and evaluation of water quality data and another project funded by NSF on precipitation-runoff modeling. Dr. Harmancioglu's basic areas of research are water resources planning and management, simulation of hydrologic processes, design of hydrologic data collection systems, erosion control, environmental data management, and information theory as applied to water resources. Currently, she is involved with studies on design of water quality monitoring networks, GIS applications in water and land resources, hydrometric data banks, identification of non-point source pollution, and watershed modeling. Dr. Harmancioglu conducted various research projects funded by DEU, TUBITAK (The Scientific and Technical Research Council of Turkey), DPT (State Planning Agency), NATO, British Council, and IWMI (International Water Management Institute). She is currently a member of such international organizations as AGU, AWRA, EWRA, IAEH, IAHS, and IWA, NYAS. She is a life-long fellow member of IAH and is currently acting as a consultant to DSI (the State Hydraulic Works) of Turkey.

Vijay P. Singh was born on July 15, 1946, in Agra, India. He obtained a B.S. in Engineering and Technology in 1967 from Pant College of Technology in India; an M.S. in Engineering specializing in Hydrology in 1970 from University of Guelph, Ontario, Canada; a Ph. D. in Civil Engineering with an emphasis on Hydrology and Water Resources in 1974 from Colorado State University; and a D. Sc. in Engineering in 1998 from the University of Witwatersrand, Johannesburg, South Africa. He is a registered professional engineer and a registered professional hydrologist. Currently, Professor Singh holds the Arthur K. Barton Endowed Professorship in Civil and Environmental Engineering at Louisiana State University. He has received more than 30 awards, including the Distinguished Service Award from the National Research Council of Italy in 1995; the Fulbright Scholar Award in 1997; International Man of the Year Award from the International Biographical Center in 1997; the Brij Mohan Distinguished Professor Award in 1999; the Distinguished Faculty Award in 1999; the Achievement in Academia Award in 1999 from Colorado State University College of Engineering; James M. Todd Technological Achievement in 2000 from Louisiana Engineering Society etc. He is a fellow of ASCE, AWRA, IE, IAH, ISAE, and IWRS. He has authored 9 text books, edited 25 books. He is Editor-in-Chief of Water Science and Technology Library Book Series and is a member of 9 journal editorial boards. Professor Singh serves as Senior Vice President of American Institute of Hydrology, Vice President of Indian Association of Hydrologists, and President of G. B. S. Board. Professor Singh's research interests have encompassed a wide range of topics in both surface and subsurface water hydrology, watershed hydraulics, irrigation engineering, and water quality engineering. He has extensively worked on kinematic wave modeling; hydrodynamics of surface irrigation; erosion and sediment transport in upland watersheds; point and non-point source water quality modeling; hydrologic modeling of ungaged watersheds; flow forecasting; areal rainfall; dam break modeling; parameter estimation for frequency distributions; multivariate stochastic analysis of hydrologic extremes; entropy modeling in hydrology; network design; landfill hydrology;

saltwater intrusion in coastal aquifers and ground water modeling. Professor Singh is also actively involved in charitable activities. He founded the G. B. School in 1994 in Agra, India. The school imparts quality education to children in rural India. He recently founded the Foundation for the Aggrandizement of Rural Areas (FARA).