Control No. APSD-87-13

AN EXPERT SYSTEM/INTELLIGENT INTERFACE FOR ACID RAIN ANALYSIS

by

A.S. Fraser, D.A. Swayne, J. Storey
and D.C.L. Lam

National Water Research Institute
Canada Centre for Inland Waters
Burlington, Ontario, Canada L7R 4A6
NWRI Contribution #87-36

MANAGEMENT PERSPECTIVE

The following work is prepared under study APSD 508 and contains relevant information from DSS contract No. KW405-6-0115, "Implementation of an expert system for the regional analysis of acid rain". The paper is prepared from a computing perspective intended for a computer science readership. The work documented in this paper details the structure and interfacing capability in the design and development of the RAISON-MICRO facility. This system is now becoming a powerful tool in the analysis of the risk assessment to aquatic resources under the LRTAP program as well as providing a platform for the evaluation of other data sets and models in various applications. The strength of the system lies in the capability to manipulate both temporal and spatial data into various configurations at the interactive control of the investigator.

PERSPECTIVE DE GESTION

Le travail qui suit est réalisé dans le cadre de l'étude APSD 508 et contient de l'information pertinente tirée du contrat KW405-6-0115 du MAS, "Mise en oeuvre d'un système expert pour l'analyse régionale des pluies acides". Le document, rédigé dans une perspective informatique, est destiné à un public d'informaticiens. Les travaux présentés ici décrivent en détail la structure et la capacité d'interface qui entrent dans la conception et le développement de l'installation RAISON-MICRO. Ce système devient un puissant outil d'analyse pour l'évaluation, dans le cadre du programme TADPA, des risques auxquels sont exposées les ressources aquatiques, et constitue aussi une base permettant d'évaluer d'autres séries de données et modèles dans diverses applications. L'intérêt du système réside dans la possibilité de manipuler à la fois des données temporelles et des données spatiales pour en tirer diverses configurations, grâce au contrôle interactif exercé par le chercheur.

ABSTRACT

This paper describes the development (current and proposed) of a user workstation for the analysis of acid rain data. Data from several large governmental collections are uploaded to an IBM PC/AT. These data cover watershed aquatic chemistry, sensitivity to acidity, volume of water discharge from sub-regions, acid deposition, and various geographical parameters. The query language is map-based, and the analysis package uses a special-purpose spreadsheet. Several functions which are currently in use as discriminants in assessing the state of a particular watershed are built-in. Feedback from the analysis is via map coloring returned to the map subsystem through a built-in spreadsheet function. Various other graphical data exploration tools are provided.

The purpose of the system is to allow the user to manipulate large amounts of data, develop and test models, and calibrate these models in regions where the effects of acid rain are documented. The experiments, once tested on well-understood watersheds or regions, may be transferred to other geographical areas.

RÉSUMÉ

L'étude décrit le développement (atteint et prévu) d'un poste de travail pour l'analyse des données sur les pluies acides. Les données provenant de diverses archives gouvernementales sont téléchargées sur un IBM PC/AT. Ces données couvrent la chimie de l'eau dans un bassin versant, la sensibilité à l'acidité, le volume de l'écoulement d'eau dans les sous-régions, le dépôt acide et divers paramètres géographiques. Le langage d'interrogation est à base cartographique, et le logiciel d'analyse comporte un tableur spécialement conçu; il est aussi doté de plusieurs fonctions qui servant actuellement de discriminants pour évaluer l'état d'un bassin donné. Les résultats de l'analyse apparaissent sous la forme de mise en couleur de la carte par le sous-système cartographique grâce à une fonction intégrée au tableur. Le poste offre divers autres outils graphiques d'exploration des données.

Le système a pour objet de permettre à l'utilisateur de manipuler de grandes quantités de données, de développer et de tester des modèles, et d'étalonner ces modèles dans les régions où les effets des pluies acides sont bien connus. Les expériences, une fois testées sur des régions ou des bassins connus, peuvent être appliquées à d'autres zones géographiques.

An Expert System/Intelligent Interface for Acid Rain Analysis

A. S. Fraser[1]

D. A. Swayne[2]

John Storey[2]

and

D. C.-L. Lam[1]

[1]National Water Research Institute,
867 Lakeshore Road, Box 5050,
Burlington, Ontario, Canada, L7R4A6
and
[2]Department of Computing and Information Science,
University of Guelph, Guelph, Ontario, Canada, N1G2W1.

INTRODUCTION

For the past two years, a collaborative effort between the University of Guelph Department of Computing and Information Science (CIS Department) and the National Water Research Institute (NWRI) has led to the development of a sophisticated user workstation and expert system for the regional analysis of acid rain data.

The user workstation development timetable has been driven by the needs of the investigators at NWRI, and much of the workstation is at an advanced state of development. The expert system function has also progressed, due to the high degree of interaction between the suppliers (from Guelph CIS) and the users (from NWRI).

The RAISON Micro system (Swayne et al. 1986) is designed to examine the relationships between terrain sensitivity indexes which assess susceptibility to acid deposition according to geologic and soil factors, and resultant aquatic chemistry. Interactive procedures making use of geographic indicators for region, watershed, and sampling site permit fast and concise data retrieval and data base editing. External computer systems including mainframes can be accessed for data transfer via menu driven protocols. Statistical analysis including frequency distributions and step-wise multiple regression analysis as well as mean, median, standard deviation and standard error can be produced on any of the data variables through a menu driven analysis package. The system is also able to provide canonical data from specified watersheds to a programmable function library, for heuristic comparisons.

Applications of these system attributes have led to the definition of certain model coefficients which are useful in the assessment of water resources at risk in eastern Canada due to the long range transport of atmospheric pollutants. Derived functional relationships can then be applied to watersheds where geological data exist but no water

chemistry has been acquired, thereby extending the geographic range within acceptable confidence limits.

WORKSTATION ENVIRONMENT

The physical workstation consists of an IBM PC/AT with two screens (one monochrome and one medium-resolution or extended graphics), at least 512K of memory, a 20 MB hard disk, an auto-dial modem and a mouse. Support software consists of the usual text editor packages, statistical packages, the Lattice C[1] compiler, the Multihalo graphics package from Lattice, and various debugging and tracing tools.

The data sources for the system reside on a mainframe computer at NWRI, in several large Environment Canada data collections. Software consists of three main systems. The first and most critical of these is the DBMS (database management system). The second is the spreadsheet and programmable function library. The third is the map subsystem.

DBMS SUBSYSTEM

The DBMS subsystem is a rather large collection of programs written in C, together with a number of index files, a database package which uses a relational data model (for time-independent or summary data about physical, chemical, or derived properties of watersheds), and a home-grown version of a balanced binary (B-tree) indexed storage scheme for the time-series data collected at numerous water chemistry and water flow monitoring stations contained within each of many of the watersheds. See for example (Wiederhold 1983).

The data collections which are sampling station-based are grouped by watershed, then by regions (aggregates of watersheds) and finally by political domain (e.g. Quebec). This hierarchical arrangement is entirely flexible - in another jurisdiction, more or fewer levels might be appropriate. Water quality and water volume measurements may not exactly correspond in time or in space, so synonyms and time-windows are applied to the water quality data. Time-independent data is made available to the time-series analyses, but only summary data from time-series may be exported to the spatial domain (as, for example, the result of a time-series calculation).

The user is free to invent new data descriptions with the aid of a highly interactive "data window". The source of this data may be either an existing data collecion or the results from a previous calculation.

---

[1]Multihalo and the Lattice C version of the C programming language written for the IBM Personal Computer are distributed by Lifeboat Associates.

SPREADSHEET

The second subsystem consists of a spreadsheet with a considerable and growing programming language interpreter and function library. The standard mathematical functions, some special functions defined for water quality, and string and list manipulation are built into this interpreter.

The spreadsheet cells and cell ranges form the objects in the language, which may contain data of type real, string, function specification, or integer. Cell ranges which have associated data names which are chemical species such as Na (Sodium), etc., have the additional property that they inherit the units of measure by which they were recorded from the database where they originated. Other permissible units, such as those expressed in gram-molecular weight are automatically attached to the cell range for a particular chemical species, so that the originator of a calculation involving mass or gram-molecular units need not worry about the units currently in force in the spreadsheet cells.

There are limitations on the use of cells or cell ranges when units or functions are defined. The dual nature of a spreadsheet cell (function or data) together with a general principal that the first-named cell in a range is the owner of the properties of the range has so far not led to difficulty.

The spreadsheet has a number of functions built in for the display of data in graphical form. Columns may be selected pairwise for x-y plotting. So-called **ion rosettes** (Fraser 1986) and other special-purpose graphical tools are available in the spreadsheet menu. The spreadsheet menu is hierarchical in nature. All of the graphical tools, for example, are contained in a sub-menu. The help menu displays the complete inventory in a convenient drop-down window invoked by a function keystroke.

A data dictionary manipulation is also contained within the spreadsheet. The "data window"program mentioned in the previous section is used to create descriptions of the available data items in the system. Using it, any number of schemata or data descriptions for a particular study area, watershed or water quality monitoring station are displayed together, and a choice of data elements to be loaded into the spreadsheet columns may be made from the individual databases. The choice is made through moving the cursor by mouse or arrow keys and selection is affirmed by highlighting. So-called **critical items** may be earmarked. Data which is missing is censored on the critical items.

The identification of the chemical species (previously mentioned) is made at this time, to provide identification of the correct gram-molecular units of measure.

User commands typed into the system are stored in a history file of previously specified length. This facility allows replication of an operation which has been recorded (similar to the use of history files in Unix). Sequences of commands for the spreadsheet may be built and interpreted by the command language interpreter which replaces menu

selection and cursor movement with operator keywords to allow a considerable latitude for programmable functions.

MAP SUBSYSTEM

The map subsystem is a key feature of RAISON Micro. From a pull-down menu, the user selects one of several options as follows:

add Icon to map
spreadsheet
quit
Display Stations
Action and Snapshot
Snapshot
Print Screen

These options can be changed at will, as they are contained in a table file. Routines which match a given option need only be linked into the menu driver program.

The user may optionally explore a map representation of the region under investigation, with the option to query the name of the region, watershed or water chemistry sampling station represented on the map. The map organization parallels that of the geographical data. Political, regional, and watershed-level views of the region of study may be selected at a wide range of map scales. A "snapshot" of any map view may be saved in bit-map form. When an icon on its parent map is selected, this snapshot is selected rather than the normal system action of redrawing the particular map segment. These snapshot files are incorporated into the system and any subsequent map-based operations such as map-coloring are not compromised. Map-based functions are independent of the existence (or otherwise) of a particular snapshot.

A menu option to access data for the purpose of loading it into the spreadsheet is invoked through a transfer of control to the spreadsheet subsystem. Corresponding help files are provided in the map as in the spreadsheet subsystems.

SYSTEM OPERATION

Data is entered into the RAISON system by running a terminal emulator with a file-transfer utility, which connects with the NWRI CYBER-171 computer containing the large water chemistry, water-flow, and terrestrial databases.

The header records produced from these databases are passed through a simple parser which decodes them and constructs a schema for data description of the incoming data. Samples not taken because of equipment outages or incomplete schema description are edited on input. The continued use of the data entry function has resulted in

increasing sophistication of the programs. For instance the capability for automated data retrieval from the large database has been made available, through the simple medium of redirecting input from a file previously stored in which the login and access scripts have been prepared.

Should a given option, (say change watershed), be selected, a map is drawn on the screen and either the numeric keypad or the mouse is used to move the cursor to an icon on the screen representing the desired watershed. This selection proceeds through region, watershed, and (if bottom-level analysis is required) station, where data is again selected through proximity of the cursor to the icon. The data is retrieved through reference to the map, from which the spatial orientation of the water chemistry stations, watersheds, and regions may be discerned.

Results of the analysis of various relationships such as terrain sensitivity to the potential to reduce acidity can be displayed in a graphical summary for each watershed or region and be displayed by the map subsystem by means of the color-fill function built into the spreadsheet.

The spatial orientation of the query language and the flexibility of the menu organization/retrieval function have kept this segment of the system very robust and easily modified. Changes in workstation structure often require no re-compilation of system routines.

APPLICATION

Under the Canadian Long Range Transport of Airborne Pollutants (LRTAP) program the investigation and evaluation of aquatic resources at risk due to acidification and aquatic acid stress forms the basis of this application. Eastern Canada, specifically the southern portion of the Province of Quebec is the primary study area. Quebec is divided into ten large drainage basins plus the area of Labrador. The southern regions (1 - 7) are more finely segmented into watershed areas as defined by the Water Survey of Canada, (Figure 1). These watershed areas are the base unit for computation of all aquatic chemistry and terrestrial data. The region of Southern Quebec is primarily composed of Pre-Cambrian shield granitic material to the north of the St.Lawrence River and deposits of more calcarious material to the south of the river. Bedrock geology and surficial soils are the primary indices of terrestrial sensitivity to the presence of acidic precipitation.

The water quality data base assembled for this work is a subset of the NAQUADAT system maintained by Environment Canada. Data evaluation and editing procedures have identified 53 watersheds that have sufficient data of quality to permit statistically significant relationships to be made (Fraser, 1986). Aquatic characterization of the study region to determine the framework for risk assessment evaluation allowed the selection of five major basins geographically spanning the study region and straddling the St. Lawrence River. By including areas of diverse terrestrial characteristics lying within

the primary deposition zone for acidification it is possible to draw statistical comparisons between aquatic and terrestrial data and thereby effectively prepare estimates on the magnitude and distribution of aquatic resources at risk in Eastern Canada.

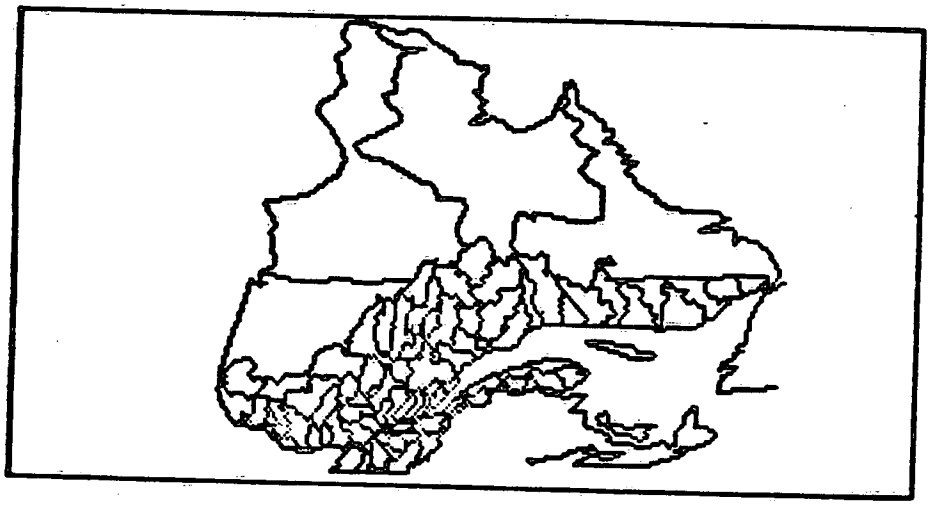

Figure 1: EASTERN CANADA SHOWING MAJOR DRAINAGE DIVISIONS AND WATERSHED AREAS IN THE PROVINCE OF QUEBEC

Consideration of the terrestrial component required adaptation of the ecodistricts defined by Environment Canada, Lands to enable overlays to be constructed which match the watershed designations in the aquatic component. The data base consists of 439 ecodistricts distributed within the 91 watersheds that comprise the total study area. A realignment and assessment of polygon overlays accomplished this task by using a geographical information system (Helie and Fraser, 1987). The first data dependent application of the RAISON map subsystem spreadsheet interaction is the evaluation of the terrestrial sensitivity data provided by Lands Directorate, Environment Canada. Geological data consisting of bedrock type, surficial soils of derived nature, soil depth and soil texture are assessed in a functional decision tree structure of hierarchical significance (Lucas and Cowell 1984). The resulting factors produce three classes (high, medium, low) potential to reduce acidity present or deposited upon a watershed. A numerical scale is then applied to the three classes to allow numerical analysis to proceed. There exist large differences in scale and resolution between the terrestrial data set which is constructed on small scale areas defined as ecodistricts and the larger sized watersheds used in the RAISON system. To overcome this problem the percentages of each numerical factor class at the ecodistrict level are area-weighted and summed for each ecodistrict within a watershed followed by the area weighting and accumulation of all ecodistricts the resolution of the individual watersheds.

The resultant numerical watershed sensitivity factors are stored in a spreadsheet file and appended to the data base. By means of the programmable function library relational functions assigning color to a variable numerical scale are developed to best identify the differing levels of sensitivity to acid deposition. Figure 2 is a direct screen dump

in grey scale showing the results of this analysis.

A comparison between the RAISON system representation of terrestrial sensitivity to acid precipitation is in good agreement with the much higher resolution cartographic representation (Li 1985). Subsequent analyses will investigate frequency class ranking for the aquatic data base and estimated of data based comparison between the terrestrial and the aquatic data by assigning color functions to such terms as correlation coefficient, absolute error, relative error, and significance testing.

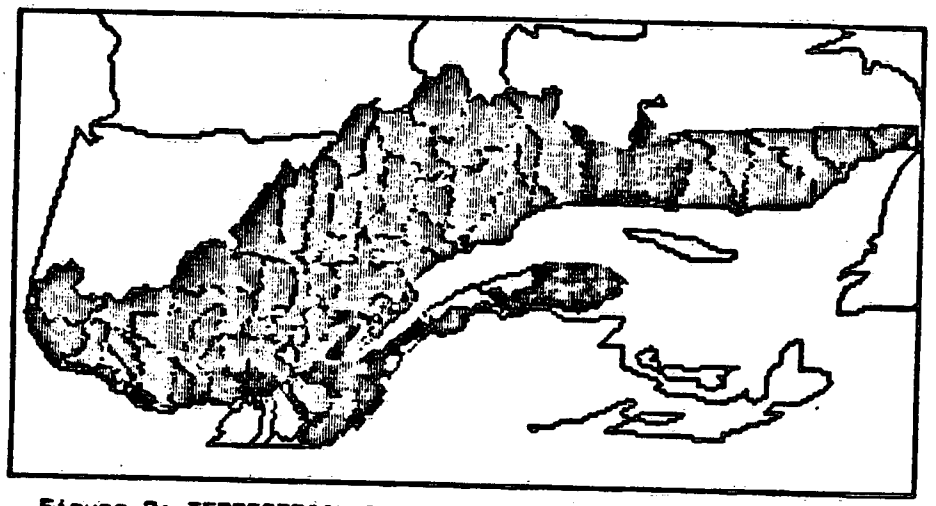

Figure 2: TERRESTRIAL SENSITIVITY TO ACID DEPOSITION FOR SOUTHERN QUEBEC AS DEPICTED BY MICRO-RAISON

WHITE = LOW BLACK = MEDIUM GRAY = HIGH

Current work is centered on the development of a knowledge acquisition facility for the interpretation of differences between model-based color functions at the watershed level. Knowledge concerning agreement (and disagreement) between model predictions and actual acidity levels form the basis for the first tentative steps towards true "expert system" functionality. This work is in process at the time of writing, and will be thus for much of the current year's development.

SUMMARY AND CONCLUSIONS

This paper has provided an overview of a workstation which supports a very specific custom approach to the analysis of acid rain data. It has a rich set of special functions, and hooks have been put in place to enhance its performance via the formation of rules and the joining of appropriate functions applied previously in a satisfactory fashion to similar data elsewhere in the database. The very uneven character of the raw data and its relative abundance has led to this "top-end" rather than a "front-end" processor, with considerable and growing ability to censor and measure the quality of the several large databases involved and to record and replicate the analysis techniques of expert investigators.

The immediate goal ahead for the RAISON development project is to allow estimates to be made of the aquatic resources at risk from acid deposition in Eastern Canada. System characteristics are presently in a steady state where this task may be considered short term. The medium term program will consist of establishing regional aquatic risk scenarios based upon aquatic/terrestrial data based relationships for other areas of Canada such as the Maritimes, Ontario, and the West coast. Lastly, the long term outlook is to begin a rudimentary pattern matching analysis system driven by an artificial intelligence protocol. This latter development is aiming to provide automated analysis procedures, building upon the experience of the users of the workstation, with the capability of running independently with minimal user intervention (unless desired). This part of RAISON is intended to assist the interpretive phase of the investigations, and eventually to complement the expertise of the investigator.

ACKNOWLEDGEMENTS

Support for this work from Environment Canada through (Canadian) Department of Supply and Services Contracts 03SE.KW405-5-0785, 02SE.KW405-5-1632, 02SE.KW405-5-2271, and 09SE.KW405-6-0115 is gratefully acknowledged.

Research assistants W. Hunt, K. Brown, T. Emburson, P. Wong, L. White, and S. White of the University of Guelph contributed substantially to the development of the workstation herein described.

REFERENCES

Fraser, A. S. "Aquatic characterization for resources at risk in Eastern Canada." Water, Air, and Soil Pollution 31, (1069-1078), D. Ridel. 1986.

Helie, R. G. and Fraser, A. S. "An integrated ecological data base in support of predictive surface water acidification modelling." Proc. Perspectives on land modelling, Lands Dir. Environment Canada. 1987.

Li, L. K. "Acid Precipitation Sensitivity Evaluation of Quebec." ELC series #20, Environment Canada, 1985.

Lucas, A. E. and Cowell, D. W., "Regional assessment of sensitivity to acidic deposition for eastern Canada." In *Geological Aspects of Acid Deposition,* Butterworth, Toronto, 1984, Ch.6 (113-139).

Swayne, D. A., Fraser, A. S., Lam, D. C.-L. and White, L., "Micro-computer analysis of aquatic effects due to acid precipitation: Part I, development of user software." In *Proceedings of the 1st Canadian Conference on the Applications of Micro-Computers in Civil Engineering* (McMaster University, May 21-22, 1986), 15-21.

Wiederhold, Gio. 1983. *Database Design,* (2nd Ed.), McGraw-Hill.