NM WAIDS: A Produced Water Quality and Infrastructure GIS Database for New Mexico Oil Producers

(Third Semi-Annual Technical Progress Report)

Report Start Date: March 1, 2003
Report End Date: September 1, 2003

Martha Cather, PI
Robert Lee, Project Manager
Ibrahim Gundiler, Co-PI
Andrew Sung, Co-PI

Contributors:
Naomi Davidson
Ajeet Kumar Reddy
Mingzhen Wei

September 24, 2003

DE-FC26-02NT15134

New Mexico Petroleum Recovery Research Center
New Mexico Institute of Mining and Technology
Socorro, NM 87801
(505) 835-5142

DISCLAIMER

"This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, produce, or process disclosed or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or any agency thereof. The views and opinions of authors expressed herein do not necessarily stat or reflect those of the United States Government or any agency thereof."

ABSTRACT

The New Mexico Water and Infrastructure Data System (NM WAIDS) seeks to alleviate a number of produced water-related issues in southeast New Mexico. The project calls for the design and implementation of a Geographical Information System (GIS) and integral tools that will provide operators and regulators with necessary data and useful information to help them make management and regulatory decisions.

The major components of this system are:

1. Databases on produced water quality, cultural and groundwater data, oil pipeline and infrastructure data, and corrosion information.
2. A web site capable of displaying produced water and infrastructure data in a GIS or accessing some of the data by text-based queries.
3. A fuzzy logic-based, site risk assessment tool that can be used to assess the seriousness of a spill of produced water.
4. A corrosion management toolkit that will provide operators with data and information on produced waters that will aid them in deciding how to address corrosion issues. The various parts of NM WAIDS will be integrated into a website with a user-friendly interface that will provide access to previously difficult-to-obtain data and information.

Primary attention during the first six months of this project was focused on creating the water quality databases for produced water and surface water, along with collecting of corrosion information and building parts of the corrosion toolkit.

Work on the project to date includes:

1. Creation of a water quality database for produced water analyses. The database was compiled from a variety of sources and currently has over 7000 entries for New Mexico.
2. Creation of a web-based data entry system for the water quality database. This system allows a user to view, enter, or edit data from a web page rather than having to directly access the database.
3. Creation of a semi-automated data capturing system for use with standard water quality analysis forms. This system improves the accuracy and speed of water quality data entry.
4. Acquisition of ground water data from the New Mexico State Engineer's office, including chloride content and TDS (Total Dissolved Solids) for over 30,000 data points in southeast New Mexico.
5. Creation of a web-based scale prediction tool, again with a web-based interface, that uses two common scaling indices to predict the likelihood of scaling. This prediction tool can either run from user input data, or the user can select samples from the water analysis database.
6. Creation of depth-to-groundwater maps for the study area.
7. Analysis of water quality data by formation.

8. Continuation of efforts to collect produced water quality information from operators in the southeast New Mexico area.
9. Qualitative assessment of produced water from various formations regarding corrosivity.
10. Efforts at corrosion education in the region through operator visits.

Future work on this project will include:

1. Development of an integrated web and GIS interface for all the information collected in this effort.
2. Continued development of a fuzzy logic spill risk assessment tool that was initially developed prior to this project. Improvements will include addition of parameters found to be significant in determining the impact of a brine spill at a specific site.
3. Compilation of both hard copy and online corrosion toolkit material.

TABLE OF CONTENTS

DISCLAIMER 2

ABSTRACT 3

LIST OF FIGURES AND TABLES 6

INTRODUCTION 7

EXPERIMENTAL METHODS 7

PROGRESS AND DISCUSSION OF RESULTS 7

Database Construction 7

Corrosion Management Toolkit 15

Formation Water Analysis 16

CONCLUSIONS 18

FIGURES 20

TABLES 26

REFERENCES 29

LIST OF ACRONYMS AND ABBREVIATIONS 30

APPENDIX 1. COMPILATION OF RESULTS FOR OPERATOR VISITS 31

LIST OF FIGURES AND TABLES

Fig. 1. Flowchart of the new string comparison algorithm 20

Fig. 2. Water analysis form. 21

Fig. 3. Reported and calculated values of chloride and magnesium in ppm and mg/l vs epm units. 22

Fig. 4. Example of Stueber's water chemistry groupings 23

Fig. 5. Location of three cross sections and 10 km buffer around them 23

Fig. 6. Cross section through Eddy and Lea counties 24

Fig. 7. Chloride concentration of produced water samples 24

Fig. 8. Cross section showing interpreted origin of the produced waters. 25

Table 1. Sample List of Data from Two Data Sets 26

Table 2. String Sample Pairs from Two Data Sets and Matching Results Sets 26

Table 3. Name and Address String Comparisons..... 27

Table 4. Conversion Factors for Common Anions and Cations..... 27

Table 5. Water Injection Statistics for Southeast New Mexico 28

INTRODUCTION

The New Mexico Water and Infrastructure Data System (NM WAIDS) seeks to alleviate a number of produced water-related issues in southeast New Mexico. The project calls for the design and implementation of a Geographical Information System (GIS) and integral tools that will provide operators and regulators with necessary data and useful information to help them make management and regulatory decisions.

The major components of this system are: 1) databases on produced water quality, groundwater data, cultural data, and corrosion information, 2) a web site capable of displaying this data in either through a GIS interface or by text-based queries, 3) a fuzzy logic-based, site risk assessment tool that can be used to assess the seriousness of a spill of produced water, and 4) a corrosion management toolkit that will provide operators with data and information on produced waters that will aid them in deciding how to address corrosion issues. The different components of NM WAIDS will be integrated into a website with a user-friendly interface that will provide access to previously difficult-to-obtain data and information.

Work in the six-month reporting period has concentrated on four major areas of the NM WAIDS project:

1. Continuation of automated data capture.
2. Data cleaning.
3. Collection of corrosion information.
4. Analysis of water data to determine regional trends in chemical variability.

EXPERIMENTAL METHODS

No experimental methods, materials, or equipment are being used for this project. Data for the water chemistry databases is being collected from previously existing private and public sources. Information on corrosion issues is being collected by mailed survey and by personal interviews with operators and chemical company representatives in the area. ArcGIS software is being used for mapping aquifers, well locations, and water chemistry data. Microsoft SQL Server 2000 is used to store and access water chemistry databases.

PROGRESS AND DISCUSSION OF RESULTS

Database Construction

A key task for this project was the development of dynamic web-accessible databases for storing, managing, accessing, and analyzing data that will eventually be accessed by both web-based queries and GIS-initiated queries. As the data files can be quite large, the system must be efficient and useable by persons with varying degrees of computer literacy. Important advances in

the database construction and web system been made. Data flows, data processing methods and user interfaces with for the NM WAIDS have all progressed.

Produced Water Quality Database

The Produced Water Quality Database (PWQD) is being compiled from a large variety of source data. A number of oil and gas producers in the area were solicited for data, and many have been very generous in sharing this information. Some of the data has been provided in digital format, either as Microsoft Excel spreadsheets, Microsoft Access databases, or simple text files. Much, however, comes from producers as paper forms supplied to them by the various companies employed to run the water analyses. Each data source must be analyzed to determine what kind of data is available and in what format (numeric, text, semi-quantitative), so the correct fields and data definitions can be built into the database structure. Examination of the digital and paper forms revealed that data could be divided into four main categories: general information, general sample properties, anions, and cations. A number of tables and views are used in the database construction: primary tables are the general sample information (items such as sample name, location, formation, physical parameters), anion information (CO_3, SO_4), and cation information (Ca, Na, Mg).

Our previous report[1] discussed the automatic data capturing system (ADCS) that has been implemented to help in processing the paper forms. In this system, water chemistry data forms are first scanned to produce a high quality image of the form. This image is then processed by optical character recognition (OCR) software to convert the image into character strings that are accepted by the database.

Although the ADCS greatly speeded the process of turning paper forms into digital data, it introduced several complications of its own that required considerable work to overcome. Two major issues had to be addressed. One was that of data conversion quality, eg., how accurate was the OCR process for recognizing numerical data? The second was that of data standardization and assimilation. Both of these problems have been addressed in this report.

Data Standardization and Assimilation

Data standardization and assimilation was the larger of these tasks. In order to assimilate data into a preexisting data structure that will allow for the construction of meaningful search queries, (such as search by location or search by well name) the data has to exist in a standardized format. Then, when the data is incorporated into the database, there must be an assurance that it does not duplicate data that is already in the database.

Well location and well name are critical in our data schema. With a correct name and location, it is generally possible to link data in the produced water database to other databases maintained by both the state of New Mexico and the Petroleum Recovery Research Center that contain information such as API, location in X,Y coordinates, correct well and operator name, field, pool, and production information for most of the wells in the state. It is highly desirable that our own data on water be easily cross-referenced to all this existent data.

When water quality forms are scanned, the OCR process simply turns the pixels in a particular region of the image into a text string. The operator of the system must then specify to

the software program that a certain region on a particular form contains information of a particular type; location, for example. Then the program places the text string it finds in that region into the appropriate location field in a preliminary database. The problem for our example field, location, is that well locations can be designated in many different ways on these forms. M-11-30-7, SW/SW 11/30/7, M 11 30N 7W, 30-7-11 and M-11-30N7W are just some of the permutations that can be used to describe the same parcel of land in the southwest corner of section 11, Township 30 north and Range 7 west. Similar variations exist in well names, pool names, and field names.

Additionally, for our example of location, almost all water chemistry reports were incomplete. Locations reported in unit letter, section, township, range (ULSTR) format are not precise, but only pinpoint a well down to a 40-acre parcel of land. Most samples lacked unit letter designations, which coarsens the resolution down to 160 acres. In order to place the information into a GIS system, one of the ultimate goals of this project, samples must have an actual X,Y location in either latitude/longitude or UTM coordinates. We have another database that already contains locations in X,Y coordinates for most of the wells in the state, including most wells that were sources of produced water data. If we could link the well information in the produced water database to that in our other database, we would have location accurate to within a few hundred feet, rather than an approximation that might be several thousand feet off. This example demonstrates only one of the reasons why data standardization and cleanup became an essential part of this project. The tremendous amount of data cleanup and standardization would have been too time-consuming to do by hand. A survey of available literature on data cleaning showed that while there were some helpful algorithms already available, an improved comparison algorithm would be necessary for this task. No previous use of an algorithm exactly like the one we designed was found in our review, so we have called this a new approximate string comparison algorithm, although it does build on the work of previous researchers.

Creation of a New Approximate String Comparison Algorithm

Background

In the context of data cleaning, it is important to identify equivalent data in multiple sources, and more importantly, to identify data in multiple sources about the same real-world entity. The first problem is referred to as a field matching problem, and the second one as a record linkage problem. These two problems are independent of each other, yet mutually related. In order to identify equivalent data in the multiple sources, it is necessary to determine whether or not two syntactic values are alternative designations of the same semantic entity. For example, a method is needed to determine if "M-11-30-7," "SW/SW 11/30/7," and "M 11 30N 7W" refer to the same well. The problem can be described as a field matching problem.

The field matching problem evaluates the similarity of different values of the same attribute of an object.[2] Its solution and algorithm can be applied to solve record linkage problems. Data from heterogeneous sources often contains problems such as misspellings, typographical variations, non-standard abbreviations, or differences in the details of data format, both at field level and at record level. In database applications, string comparison algorithms for field

matching or record linkage purposes are designed to detect strings that are duplicates of each other, but not necessarily textually identical. Using the name of a person as an example, "David Smith" and "David S." could be the same person, but there are differences in characters between these two strings representing the name. Approximate string matching rather than strict character-by-character comparison is necessary for appropriate identification of these two strings as the same entity.

Approximate string comparison has been a subject of research in computer science for many years.[3] There are several approximate string comparison algorithms, each having its own characteristics. Most can be divided into one of two classes: character-by-character and word-by-word. Simply, algorithms in the character-by-character class compare strings character by character, and those in the word-by-word class compare strings word by word.

Time complexity is an important feature of any string-matching algorithm. Time complexity is defined as the way in which the number of steps required by an algorithm varies with the size of the problem it is solving.[4] Time complexity is normally expressed as an order of magnitude: O(N^2) means that if the size of the problem (N) doubles then the algorithm will take four times as many steps to complete. Some string-matching algorithms may have very good results for data comparison but the time complexity can make them unusable for all but the smallest data sets.

Character-Based String Comparison Algorithms

Character-based string comparison algorithms compare two strings character by character, although they have different strategies to handle misspellings and variations in strings. Edit distance[2,5] is a common measure of textual similarity. Given two strings, s1 and s2, their edit distance, denoted $\Delta e(s1,s2)$, is the minimum number of edit operations, insertions, deletions, and substitutions of single characters that are needed to transform s1 to s2. For example, Δe("Harrison","Harison") = 1. This algorithm compares two strings, s1 and s2, character by character and its time complexity is $O(|s1|*|s2|)$, where $|s1|$ is the length of string s1 and $|s2|$ is the length of string s2. Further refinements of this basic concept have been introduced over the years.[5–11] Refinements have included methods to allow gaps, or sequences of non-matching symbols, improvements in time efficiency, and to allow for[3,12,13] partial agreement between two strings.

Word-based String Comparison Algorithms

Word-based string comparison algorithms are different from character-by-character algorithms. These algorithms compare two strings word by word, rather than character by character. A literature search has revealed only two word-by-word string comparison algorithms thus far. One is a basic field matching algorithm and the other a recursive field matching algorithm.

The basic field matching algorithm splits strings into sequences of atomic strings and then sorts the atomic string sequences. An atomic string is a sequence of alphanumeric characters delimited by punctuation characters. Two atomic strings match if they are the same string or if one is the prefix of the other. The matching degree of two strings is defined as the number of

common atomic strings divided by the average number of atomic strings.[7] This basic algorithm will handle two situations; that where two atomic strings match each other exactly, and that where one atomic string is the prefix of the other. For example, the degree of match between the strings "Comput. Sci. & Eng. Dept." and "department of computer science" is 3*2/7 = 0.86. The algorithm processes the string matching word by word in a straightforward way, which makes the order of atomic strings in strings have no influence on the matching degree. This is the most important advantage of this algorithm. Because string pairs commonly do have some out-of-order problems, the lack of influence of order on the matching degree makes this method superior to character-based comparison algorithms. However, the algorithm still does not cover all possible string matching situations, limiting its application.

A recursive field matching algorithm[7] was developed to make up for the disadvantages of the basic field matching algorithm. The recursive algorithm improves the basic algorithm by extending the string matching situations. In addition to the two string-matching situations handled by the basic algorithm, three more are given: a) one string is the combination of a prefix and a suffix of the other string, as "dept" and "department"; b) one string is the acronym of the other string, as "UCSD" and "university of California, San Diego"; and c) one string is the concatenation of prefixes from the other string, as "CalTech" and "California Institute of Technology";

These string matching situations cover all possible string matching situations one can meet when comparing two character strings. However, this algorithm is much more time-intensive to run, and still excludes many cases encountered in real-world string-matching applications.

From the above description, one can see that there are important differences between these two classes of string comparison algorithms. Character-by-character algorithms compare strings by character so the order of characters or words in the strings is very important; word-by-word algorithms do the same thing by word, hence only atomic string-matching matters and the order of atomic strings is ignored. Word-based matching algorithms are much more flexible than character-based matching algorithms for long multiple-word strings and character-based matching algorithms are better for spelling error checking. The word-based string comparison algorithms increase the flexibility of string comparison by ignoring order problems and enrich its applications in various fields.

A New String Comparison Algorithm

Different string comparison algorithms fit different string comparison situations. Analysis of specific data characteristics is quite important prior to use of any algorithm. Analysis of our specific data problems showed data match issues that no single matching algorithm that we found was capable of handling in a time-efficient and correct manner. Table 1 shows data samples from two data sets that were used to create the final Produced Water Quality Database. This table demonstrates some of the differences between data that are equivalent for the same entity. The data on the left is from the scanned forms in a preliminary database, while the data on the left is the correct data according to the State of New Mexico's well database.

From Table 1, the following data problems can be detected.

- Out-of-order problems, like strings in pair 1;

- Typographical problems, like strings in pairs 2 and 3;
- Atomic string missing problem, like strings in pair 4;
- Abbreviation problems, like strings in pair 5;
- Numerical data matching problems, in all above samples;
- Mixture of numbers and characters in target strings in all above samples;
- Mixed atomic strings with characters and numbers, like strings in pair 6.

From the observation of the above strings, we can see three kinds of strings to match after removing all stop words and characters (words such as "and," "in," "the," "of," "inc," and special characters or punctuation marks) from strings.

- Character strings, like "SAN";
- Numerical strings, like "001";
- Mixed strings, like "2A";

Combining all these kinds of strings, the following string-matching cases can be derived from our analysis.

- Matching between character strings, like "SAN JUAN" and "SJ";
- Matching between numerical strings, like "2" and "002";
- Matching between mixed strings, like "2A" and "002A";
- Matching between a character string and a numerical string, like "April" and "04";

Based on the above string-matching analysis, a new approximate string comparison algorithm was developed specifically for our data problems. Figure 1 is a flowchart of this new string comparison algorithm. It is based on previous word-based string comparisons, particularly the work of Monge and others,[7] but is greatly expanded to include the following matching problems:

For two character strings.

- Two strings exactly match each other ignoring case, as "match" and "Match";
- One string is the prefix of the other string, as "sci" and "science";
- One string is the combination of prefix and postfix of the other string, as "dept" and "department";
- One string is the acronym of the other string, as "UCSD" and "University of California, San Diego";
- One string is the concatenation of prefixes from the other string, as "CalTech" and "California Institute of Technology";

For two numerical strings.

- Two strings match exactly if their real numerical values are same, as "012" and "12", "3.5" and "3.5000"

For two mixed strings.

- Two mixed strings match exactly if two strings both are in pattern "CharactersNumbersCharacters" or "NumbersCharactersNumbers" and the character parts of two strings match exactly according to the matching rules for character strings and the numerical parts of two strings match exactly according to the matching rule for numerical string, as "2A" and "002A".

  For one character string and one numerical string.
- Two strings match exactly if the character string is the exact alternative presentation of the numerical string, as "one" and "001".

By extending the basic field matching algorithm, our new string comparison algorithm covers all possible string matching situations found after removing stop words in strings.

Since this string comparison algorithm is used to compare strings from large data sets, time efficiency becomes a very important factor. Time complexity of the basic field-matching algorithm, which is dominated by the time involved in sorting two sequences of atomic strings, is O(nlogn) where n is the maximum number of atomic strings in either string. The time complexity of the new algorithm is same as that of the basic field-matching algorithm. The new algorithm employs an optimization method to mark all matching atomic strings to prevent matching atomic strings being revisited. This greatly improves the time efficiency of the algorithm.

We used the new approximate string comparison algorithm to compare strings representing the entity, "names of oil producers in New Mexico," in different data sets and to link corresponding records with semantically equivalent but not textually identical strings. There are more than 4,000 records to process in the target data set, and over 200,000 records in the source data set. Results of the new algorithm show a very good improvement. Table 2 shows examples from two data sets in question and matching results based on the new algorithm, compared with those using the previous character-based or word-based algorithms. Character-based algorithms are represented by an edit distance algorithm and word-based algorithms by basic field matching. From this table, it is easy to see that the new algorithm improves the degree of string matching by including more string matching situations. The new algorithm also includes a new method of calculating the degree of match.

Since approximate string-matching algorithms are widely used in name and address matches, examples are taken from these areas to test the accuracy of the new algorithm. Table 3 shows the string comparison results for some sample string pairs. From this table, one can see that the new algorithm generally gets higher matching result scores, especially for strings containing abbreviations. Examination of the two tables reveals that.

- Edit-distance algorithms are best suited for comparing two one-word strings and multi-word strings without out-of-order problems;
- Edit-distance algorithms produce poor matching results for strings having out-of-order problems and abbreviations;
- Basic field matching is suitable for comparing two strings having all atomic strings being character strings, and numerical strings with the same format and scale;
- The new algorithm produces good results for strings with character strings, numerical strings and mixed strings;

- Both the basic field-matching algorithm and the new algorithm have difficulty with strings containing spelling problems, especially for one-word strings with spelling errors. The new algorithm appears to handle multi-word strings with spelling errors more accurately than the other algorithms.

Even with application of this improved string-matching algorithm, ambiguous cases were found where the computer alone could not make the link between data from one source and another. These cases were all flagged and examined by hand. In many cases, it was possible to correctly link the data. There were, however, a number of instances where there was simply not enough information associated with the water quality report to accurately link the water data to important identifying information from our other databases. Such an example would be a water report that identified the sample source as State Com #1, or Sample #23, or Tank Battery SJ 30-6. These cases will not be included in the final database.

Data Quality Control

One additional phase of data quality control was that of checking the accuracy of the ADC system. Although the operators of the system tried to be conscious of checking data accuracy, it was felt that an additional check on data quality was necessary, particular for the numeric water chemistry data. One of the best ways to check data accuracy was to examine the relationships in major element chemistry. Most water sample reports have data presented in two different units (Fig. 2). This might be in parts per million (ppm), milligrams per liter (mg/l), milligram equivalents per liter (me/l) or equivalents per million (epm). Equivalents per million, a unit of measurement involving the number of ions, is often used in studies of chemistry of natural waters and in the interpretation of analyses. In waters of low salinity, the unit epm is numerically the same as the unit milligram equivalents per liter (me/l). For practical purposes, they can be considered identical. Concentrations expressed in units of weight as parts per million (ppm) are sometimes desired for a particular purpose [14] and are a very common reporting unit. Conversion factors for some common anions and cations are included in Table 4. The equivalent weight of chloride is 35.5; thus 5 epm of chloride is the same as 177.5 ppm and 1 ppm chloride = 0.0282 epm.. These linear relationships can be used to check data quality. Figure 2 shows graphs of chloride and magnesium reported in ppm vs epm for some of the scanned data; similar checks can be run on all other data.

Figure 2 shows that most reported data points lie on or very close to a line whose slope corresponds to the conversion values determined by ppm/epm. Points that lie significantly away from this line are considered suspect. These data points are flagged and checked against the scanned images of the data forms. In approximately half the cases, the error was found to be in the conversion of the image to text. The most common conversion error occurred in cases where the OCR program could not distinguish between a comma and a decimal point. In the other half of the cases checked, the OCR conversion was correct, and the problem lies in the actual data itself. A decision was made to keep the data in the database, but to add the notation to that particular data field that the data itself may have an error. In this way, the decision to use the data is left up to the individual database user.

At this point the Produced Water Quality Database is nearly complete. A minor amount of cleanup remains to be performed, and there is still more data being added; however the database

is ready to be used. The entire database contains approximately 7500 records. Of these, about 1500 are from Lea, Eddy, and Chaves counties. The remainder comes primarily from the San Juan Basin and southern Colorado. Although these data are from wells outside the study area, strong industry support compelled us to add this data into the database.

Corrosion Management Toolkit

A second key area of the NM WAIDS project is the Corrosion Management Toolkit (CMT). It has been determined that the toolkit will consist of the following components:

1. Reference Book: Descriptions of common types of corrosion-related problems, including photos or drawings. This portion will also include possible mitigation approaches. The web version of this component may also include contact information and links to companies that provide these solutions in the Permian Basin area.
2. Formation Water Analysis: A qualitative analysis of produced waters in the region, by formation, with respect to corrosion related problems. Regional variations and trouble spots will be noted, and if enough data becomes available, we will generate maps to graphically depict the variations.
3. Scale Prediction Tools: Online tools that will allow the user to predict scaling tendencies of either a single water or a mix of waters. The tools will allow the user to input their own data, or they can request samples from the PWQD if they lack sample data. A simple online corrosion prediction tool may be provided. However, this is a very complex and not easily-quantified problem; many chemical vendors have spent considerable time and effort developing commercial software products to do this.

Progress

Significant progress has been made in the construction of the Corrosion Management Toolkit. The framework of the web site has been constructed and placed on a development web server. The web site will enable users to link to both produced and groundwater databases, the various prediction tools, and online corrosion manual. Formation water data has been analyzed for both general geochemical trends and regional variations. Regional oil and gas well operators and chemical providers have been contacted and interviewed regarding specific corrosion-related issues.

Web Pages

Web page construction was begun during this reporting period. Several parts of the web have been completed: for example, the scale calculation tool, database queries, and parts of the GIS maps. The next step will be to place all of these components on a development server and integrate them into a single web site. Primary concerns now are to ensure that the site is user-friendly and works in all commonly used web browsers (Internet Explorer, Netscape, Mozilla). Currently some applications work well in some browsers but not others. A secondary concern, which will be addressed during the next reporting period, will be to create the web-based map

interface using ArcIMS.

Reference Book and Problem Assessment

We have intensified our efforts to contact operators and chemical companies for information regarding corrosivity of specific formation waters. In March 2003, a letter was sent to approximately 100 service companies and chemical vendors who operate in the Permian Basin. In this letter, workers familiar with the area were asked to fill out a form with information about various formations. Very few companies actually responded to this survey, so we started the process of making personal contacts with companies, asking for information and cooperation. In most cases, we provided a map of the Permian Basin, and asked our contacts to note on the map specific regions where they found problems and what those problems might be. This method has produced promising results and will be continued during the next reporting period. Results to date are included in Appendix 1. The most common problems thus far seem to be associated with H_2S, which is reported by almost all who were interviewed.

More operators will be interviewed during the next three months, and additional details about problem areas will be collected. The next step for reference book compilation will be to focus on the problems that have been identified thus far. For these problems, we will obtain photos and diagrams that help with identification, suggest a variety of mitigation methods and include some discussion of the methods that have been tried to date, and finally, provide contact information for regional providers who may be able to help or who have expertise in this area.

Formation Water Analysis

Water injection

Because of the large number of injection wells in southeast New Mexico, there was a question as to whether a study of produced water chemistry was even valid. Water injection occurs over the entire study area, (Fig. 5) in nearly all formations, with the exception of those formations with low porosities and permeabilities ("tight" formations). Two types of water injection wells are found in the area. A simple water injection well is used to keep pressure up within a formation, and usually injects water of similar composition to the water produced from that formation. A salt-water disposal well (SWD) is a deep well used to dispose of large amounts of highly saline brines, which may be of any composition. Most SWD wells inject into highly permeable formations that are thought to be well isolated from either hydrocarbon-producing zones or potential aquifers.

Water injection for pressure maintenance typically uses produced water from the same formation in order to prevent scaling within the formation from mixing incompatible waters. Nearly the same amount of water is injected that is produced (see Table 5); however, the composition of the injected water is usually somewhat different from the original produced water composition. There are several reasons for this difference in composition. The first is that there is often mixing of produced waters in a holding tank within an oil field; as producing and reinjection continues over time, the composition of water within a formation will tend to homogenize over distance. The second reason is that oil makes up a certain percentage of the produced water, and

for injection into the ground, the lost volume of liquid is often made up by fresh water (often from the Ogallala Formation.) Again, this is to prevent possible scaling from mixing two incompatible brines; water from the Ogallala is fresh enough that the produced brine will only be diluted. Although the large amounts of water that are produced and re-injected does mean that water cycling occurs (in which the water in a formation is replaced over time by injected water), the water remains compositionally similar to the original water. In addition, the study area is so large that regional trends will not be affected much; thus, it is valid to use the produced water to evaluate whether and where meteoric waters are moving in the deep basin.

Interpretation of Water Chemistry

Water chemistry analysis methods described by Steuber were used in this study.[15] In the work of Stueber, Na, Cl, and Br ions, were used to delineate samples that reflected derivation from different sources. The samples were all taken from formations composed of carbonate, in the form of calcite or dolomite. The different groups were also delineated using TDS and ratios of some major ions, including Cl^-/SO_4^{2-} and Ca^{2+}/Mg^{2+} ratios. The groups delineated by Ca/Mg ratios may or may not have any significance, because all the samples were from carbonate formations. The Cl^-/Na^++K^+ ratio in conjunction with the TDS was also examined; however, most separation between groups was believed to be a function of the TDS and there was not enough difference between the Cl^-/Na^++K^+ ratios to separate the two groups. Stueber did not include HCO_3^- in his description, so that ion is ignored; however this seems reasonable considering the highly dynamic nature of both the pH and HCO_3^- concentration when a produced water sample is taken. The analyses also demonstrated that some samples are a mixture of the two groups of water (Fig. 4).

Using the data analysis from as an example, chemistry data from the Produced Water Quality Database were grouped by producing formation into four groups[1]. The first group, as in Stueber's work, is saline meteoric water. This group has a TDS of less than 75 g/L, a Cl/SO_4 ratio of less than 50, and a Ca/Mg ratio between 2 and 4. The second group also corresponds with the second group of Stueber, the modified evaporitic marine brines, and has a TDS of 125 g/L or above, a Cl/SO_4 ratio greater than 50, and a Ca/Mg ratio between 4 and 7. The third group is a mixture of the first two groups. It has a TDS between 75 and 125 g/L, a Cl/SO_4 ratio less than 50, and a Ca/Mg ratio between 4 and 6. The fourth group includes all the samples that do not fall under the other three. Because the percentage of samples in the first three groups for each formation was often not high, it was decided to show groups based on the Cl/SO_4 ratio and TDS only (omitting the Ca/Mg ratio requirement) to augment the number of samples in each group. Once the data was grouped, the spatial variations were examined both in planimetric view using ArcView GIS and in three dimensions by plotting the data on to geologic cross sections.

Geologic Cross Sections

Three geologic cross sections were created based on water sample distribution patterns (Fig. 5). The data for the cross sections came from scout cards in the New Mexico Bureau of Geology Subsurface Library. These scout cards provided information such as well location, elevation of the well in feet above sea level, the depth of formation tops in feet from the surface,

and some DST information regarding oil production. The location, elevation, and formation tops were used to create these cross sections. An attempt was made to get good data from one well in each section that a line of the cross section crossed (i.e., one well per mile.). Most wells did not have every formation top, so some interpolation was necessary where the data was lacking.

Once the cross sections were created, all produced water samples located ten kilometers or closer to each cross section were selected using ArcGIS, and then projected onto the closest cross sections in a N-S or E-W line, depending on the predominant direction of the cross section. The samples were plotted as depth versus latitude for the N-S cross sections, or depth versus longitude for the E-W cross section. Adobe Illustrator® was used to place the water quality data on top of the cross sections.

Three sets of data were projected on each cross section, so there are three figures per cross section. Only one of these sets, the east-west cross section through Lea and Eddy counties, is shown in this report. Figure 6 shows the reported formation of the produced water sample, which illustrates how well the water sample formations match the geologic cross section. Figure 7 shows the chloride concentration of the produced water samples, for comparison with the map view figures. Chloride concentration is used as a proxy for the distribution of TDS within the cross section.

Water Analysis Interpretation

Produced water samples were obtained from almost all formations (Fig. 6). In comparison with groundwater data analyzed in a previous report, produced waters typically have a much higher average chloride concentration than the groundwater samples. The piper diagrams for all formations except those in the Artesia group are similar and show halite dissolution. The larger range of Artesia group samples in the piper diagrams is likely because of the samples with low chloride (higher water quality). Such samples are likely recent meteoric water and show the influence of the smaller amount of time for equilibration with local minerals.

Looking at the samples grouped by chemical ratio, high variability within formations and short sampling distances is evident. Thus, spatial trends are not as evident as they were for groundwater, even with large numbers of samples. However, it appears that there are more modified brines (group 2) with distance east, especially once over the Central Basin Platform, which is consistent with Dutton's[16] findings of eastward regional flow.

CONCLUSIONS

A great deal of progress has been made in all phases of the NM WAIDS project. During the past eighteen months, we have assembled large databases of both produced water quality and groundwater quality, created a web-based data entry system that allows remote users to add data to the PWQD, and created an automatic data capturing system to allow for more accurate data collection. The data has now been cleaned and is ready for use. Web-based text queries have been created to allow users to search either of these databases and retrieve information for use in the scale prediction and water mixing tools. We have obtained a great deal of produced water chemical data, contacted many Permian Basin operators on corrosion and other water-related

issues, and created a web-based tool to predict scale type and probability based on two of the most commonly-used industry methods. We have also begun the process of assembling qualitative water quality information that will be examined for variability by formation, region, and type of problem. Geographic coordinates have been created for all sample locations with sufficient information and we have assembled other pertinent information such as base maps for roads, metropolitan areas, political boundaries, geology, hydrology, and similar data. Aquifer thickness for the area, a possible key parameter in evaluation of brine spills, has been calculated.

Permian Basin operators have been helpful and interested in our work. Yates Petroleum Corporation of Artesia has been particularly generous in the sharing of information, time, and source codes that were a partial basis for the scale calculation tools.

Work is progressing in a timely fashion. The greatest hurdles have been in assembling and entering data for the produced and ground water quality databases, and in cleaning up these databases so that the data is as accurate as possible. Operator cooperation on specific corrosion issues has been somewhat slow to develop, but we have been increasingly successful in gaining access to their site-specific knowledge. The next few months will see the creation of the web site and assembly of the many GIS maps that have been used for water analysis. Two different fuzzy tools that have been examined by various members of the NMOCD and the NMOGA Chlorides working group will be deployed within the GIS, although neither of these tools will necessarily be the tool that is finally approved by the NMOCD and NMOGA.

FIGURES

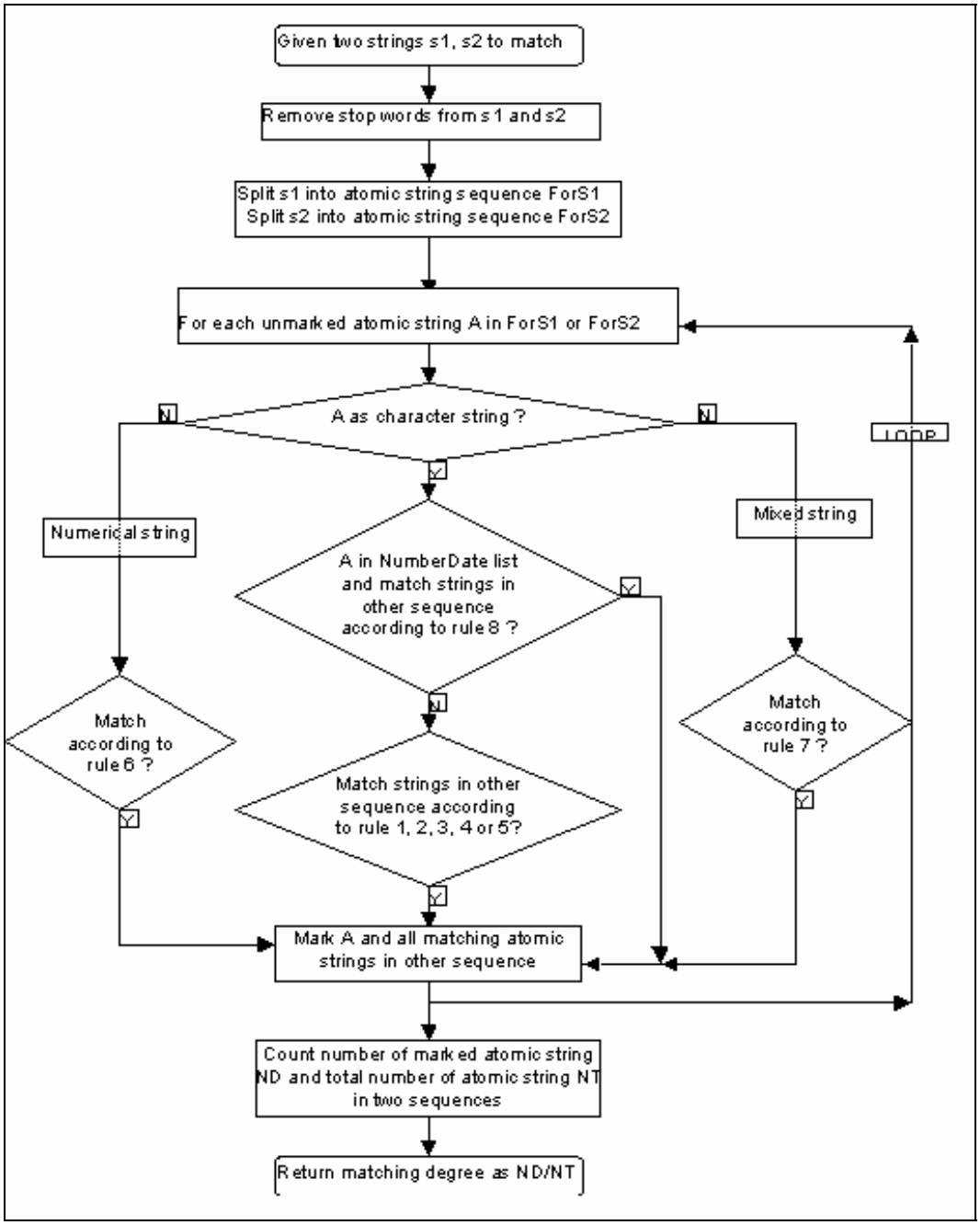


Fig. 1. Flowchart of the new string comparison algorithm used for cleaning data to be entered in the Produced Water Quality Database.

EL PASO NATURAL GAS COMPANY
SAN JUAN DIVISION
FARMINGTON, NEW MEXICO
PRODUCTION DEPARTMENT WATER ANALYSIS

Analysis No. 1-8572 Date October 01, 1975

Operator E.P.N.G. Well Name Horton #6

Location 29-75 County State

Field Formation M. V.

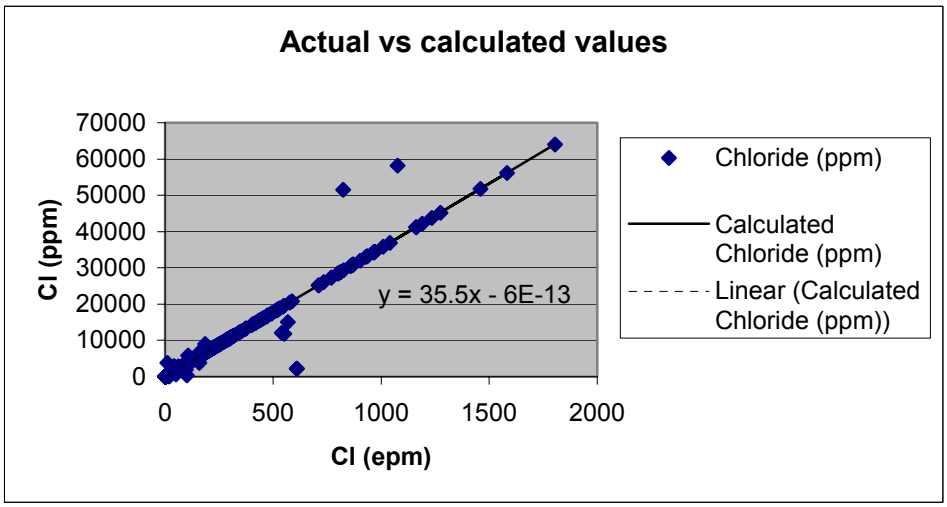
Sampled From Bradenhead

Date Sampled 09-23-75 by Thomas F. West

Tubing Pressure Casing Pressure Surface casing pressure

	ppm	epm		ppm	epm
Sodium	3330	145	Chloride	1680	47
Calcium	350	17	Bicarbonate	15	0.2
Magnesium	30	2.4	Sulfate	5615	117
Iron	Present		Carbonate	0	0
H_2S	Absent		Hydroxide	0	0

Fig. 2. Water analysis form showing two reporting methods (ppm and epm) for the sample.



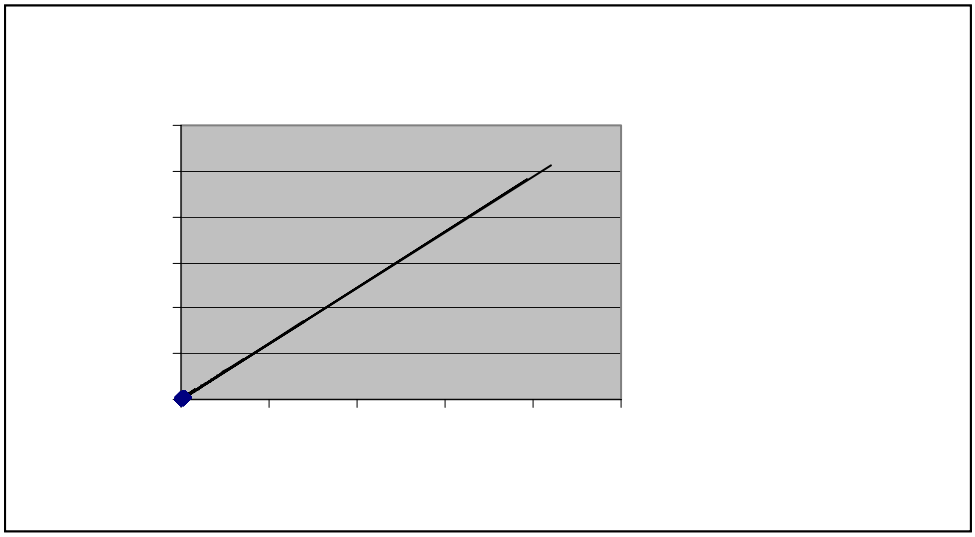


Fig. 3. Reported and calculated values of chloride and magnesium in ppm and mg/l vs epm units. These types of comparisons are very useful for quality control of scanned and hand-entered data. If the value of the number in epm units divided by that of ppm or mg/l units was less than half or greater than twice the known conversion factor (for instance, 1ppm = 0.0282 epm for chloride), the data was considered inaccurate and values were checked by hand against the scanned image. Note that the slope of the trend lines for calculated data is equal to the conversion factor given in Table 4.

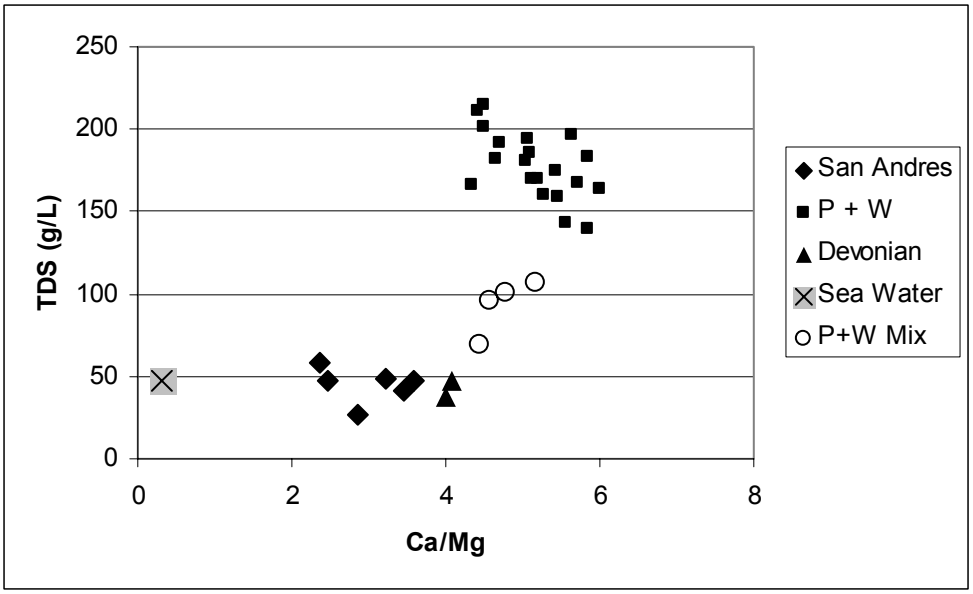


Fig. 4. Example of Stueber's water chemistry groupings. The square symbols are Pennsylvanian and Wolfcampian waters, while the open circles reflect samples with a mixed brine and meteoric water composition.

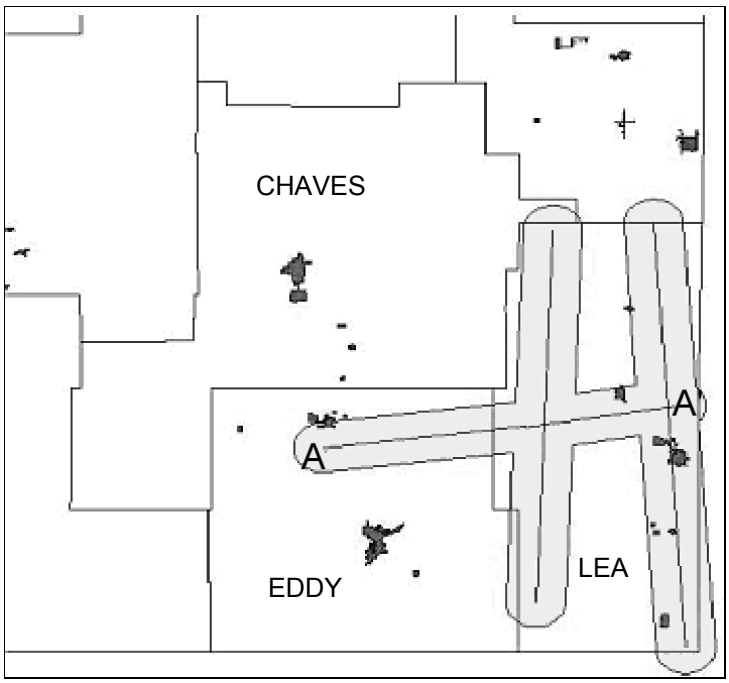


Fig. 5. Location of three cross sections and 10 km buffer around them. Produced water samples from wells that fall within the buffer zone were projected onto the closest cross sections in a N-S or E-W line, depending on the predominant direction of the cross section. Line A-A' marks the cross section line and buffer zone for the next three figures.

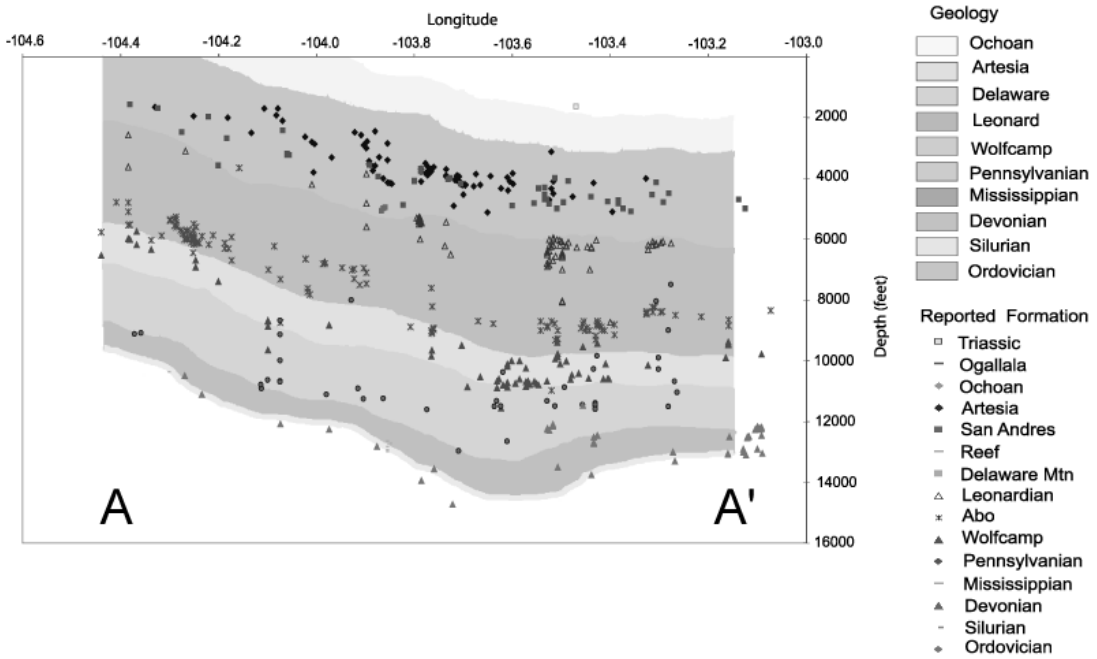


Fig. 6. Cross section through Eddy and Lea counties showing produced water samples by depth and reported formation.

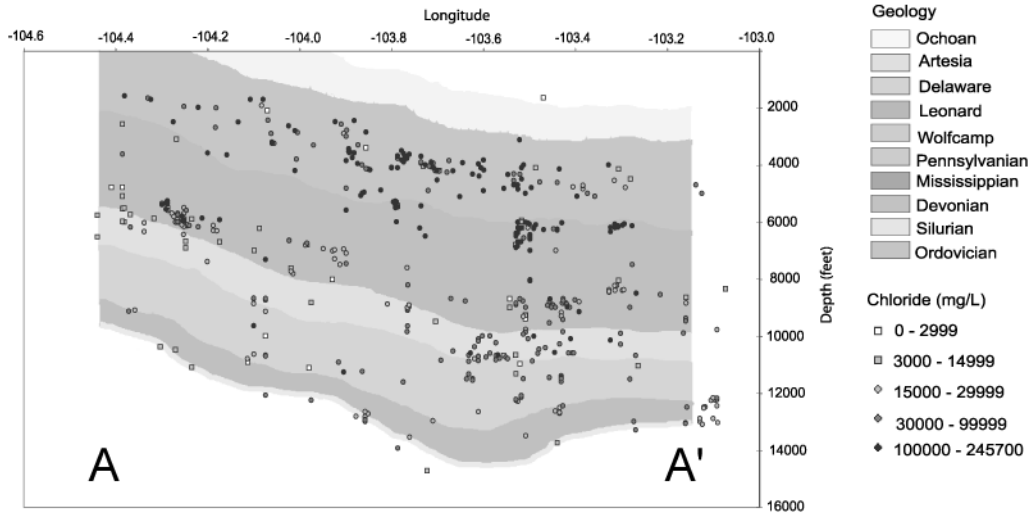


Fig. 7. Chloride concentration of produced water samples for cross section A-A'. Interestingly, some of the most saline waters are found at relatively shallow depths.

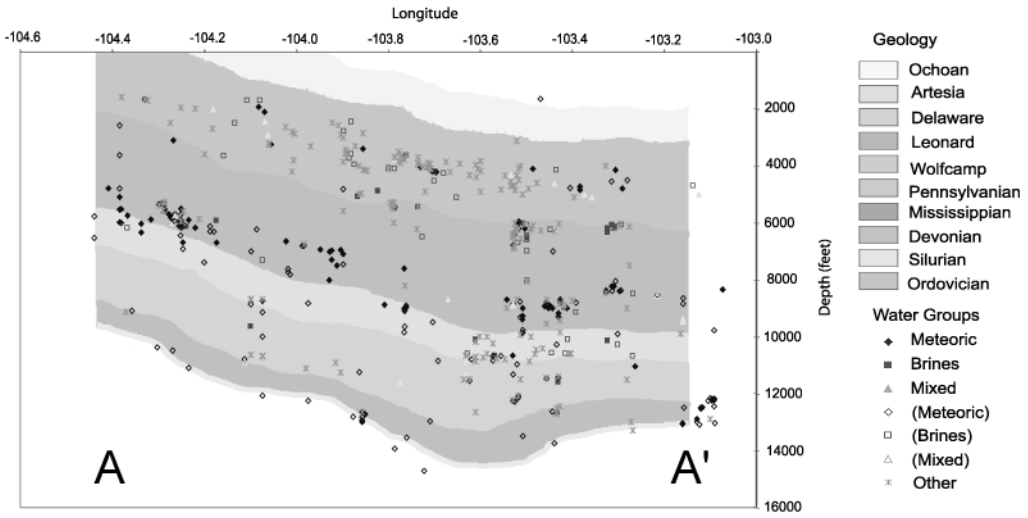


Fig. 8. Cross section A-A ', showing interpreted origin of the produced waters. Solid points are based on Cl/SO_4 and Ca/Mg ratios, while open points are based on Ca/SO_4 ratios only.

TABLES

Table 1. Sample List of Data from Two Data Sets

Pair	String From Water Quality Data Set	Matched String From State Source Data Set
1	#1 DICKINSON HEIRS	DICKINSON HEIRS 001
2	JENNING'S FEDERAL NO. 1	JENNINGS FEDERAL 001
3	FEDERAL-KEOHANE ETAL #2	KEOHANE ET AL FEDERAL 002
4	30-6 POW #3	SAN JUAN 30 6 UNIT POW 003
5	SJ 29-6 105	SAN JUAN 29 6 UNIT 105
6	GAMBLING 2A	GAMBLING 002A

Table 2. String Sample Pairs from Two Data Sets and Matching Results Sets

Pair	Strings from Water Quality Data Set	Matched Strings From Source Data Set	Edit Distance	Basic Field Matching	New Algorithm
1	#1 DICKINSON HEIRS	DICKINSON HEIRS 001	6	0.67	1.0
2	A-1 EAVES	EAVES A 001	9	0.67	1.0
3	E M HARTMAN 1	EM HARTMAN 001	3	0.57	1.0
4	C. M. FARNSWORTH B 5	FARNSWORTH B FEDERAL 005	14	0.44	0.67
5	FEDERAL-KEOHANE ETAL #2	KEOHANE ET AL FEDERAL 002	17	0.67	1.0
6	30-6 POW #3	SAN JUAN 30 6 UNIT POW 003	15	0.55	0.73
7	SJ 29-6 105	SAN JUAN 29 6 UNIT 105	12	0.8	0.9
8	GAMBLING 2A	GAMBLING 002A	2	0.5	1.0

Table 3. Name and Address String Sample Pairs and Result Comparison

Pair	Strings from Water Quality Data Set	Matched Strings From Source Data Set	Edit Distance	Basic Field Matching	New Algorithm
1	Colette Johnen	John Colette	12	1.0	1.0
2	Liu Hang Xiang	Liu H. X.	7	1.0	1.0
3	CalTech	California Institute of Technology	24	0.5	0.75
4	NM Tech	New Mexico Institute of Mining and Technology	31	0.57	0.71
5	600 113th st. apt. 5a5	600 113th st. ap. 585	2	0.6	0.8
6	SHACKLEFORD	SHACKELFORD	2	0.0	0.0

Table 4. Conversion Factors for Common Anions and Cations

Cation	Equivalent weight	Conversion Factor (1/equivalent weight)	Anion	Equivalent Weight	Conversion Factor (1/equivalent weight)
Calcium (Ca)	20	0.05	Carbonate (CO_3)	30	0.0333
Magnesium (Mg)	12.2	0.08197	Bicarbonate (HCO_3)	61	0.0164
Sodium (Na)	23	0.0435	Sulfate (SO_4)	48	0.0208
Potassium (K)	33.1	0.0302	Chloride (Cl)	35.5	0.0282
			Nitrate (NO_3)	62	0.0161
To convert epm to ppm, multiply the concentration in epm by the equivalent weight. To convert ppm to epm, divide the concentration in ppm by the equivalent weight.					

Table 5. Water Injection Statistics for Southeast New Mexico

Eddy				
		2001	2002	2003 (as of 7/21/03)
	Cum. Prod. H_2O (bbl)	172329258	166563993	64194936
	Cum. Injected H_2O (bbl)	152829075	146926205	63197886
	% of Cum. Prod. H_2O injected	89	88	98
Chaves				
		2001	2002	2003 (as of 7/21/03)
	Cum. Prod. H_2O (bbl)	7051815	6446153	2999047
	Cum. Injected H_2O (bbl)	6819553	6772747	2822281
	% of Cum. Prod. H_2O injected	97	105	94
Lea				
		2001	2002	2003 (as of 7/21/03)
	Cum. Prod. H_2O (bbl)	429351183	412605679	154455426
	Cum. Injected H_2O (bbl)	403614286	428160672	156906731
	% of Cum. Prod. H_2O injected	94	104	102
Total		2001	2002	2003 (as of 7/21/03)
	Cum. Prod. H_2O (bbl)	608732256	585615825	221649409
	Cum. Injected H_2O (bbl)	563262914	581859624	222926898
	% of Cum. Prod. H_2O injected	93	99	101

REFERENCES

1. Cather, M.E.: "NM WAIDS: A Produced Water Quality and Infrastructure GIS Database for New Mexico Oil Producers," annual report, Contract No. DE-FC26-02NT 15134, U.S. DOE, Washington, DC (March 2003).

2. Jin, L., Li, C., and Mehrotra, S.: "Efficient Record Linkage in Large Data Sets," *Proc.,* Eighth Int. Conf.on Database Systems for Advanced Applications (DASFAA), Tokyo, Japan, (2003).

3. Porter, E.H. and Winkler, W.E.: "Approximate string comparison and its effect on an advanced record linkage system," *in* Proceeding of an International Workshop and Exposition - Record Linkage Techniques, Arlington, VA, USA, (1997).

4. FOLDOC, The Free On-line Dictionary of Computing:. URL:http://www.foldoc.org/, Editor Denis Howe.

5. Navarro, G.: "A Guided Tour to Approximate String Matching," ACM Computing Surveys (2001) **33,** No.1, 31 – 88.

6. Monge, A.E.: "Adaptive Detection of Approximately Duplicate Database Records and the Database Integration Approach to Information Discovery." PhD Dissertation, University of California, San Diego, California (1997).

7. Monge, A.E., and Elkan, C.P.: "The Field Matching Problem: Algorithm and Application," *Proc.* Second International Conference on Knowledge Discovery and Data Mining, Portland Oregon, (1996), p. 267-270, AAAI Press.

8. MacCallum, R.M.: "Computational Analysis of Protein Sequence and Structure," On-line document, URL:http://www.sbc.su.se/~maccallr/thesis/thesis.html, 1997.

9. Barton, G.J.: "An Efficient Algorithm to Locate All Locally Optimal Alignments Between Two Sequences Allowing for Gaps," *Computer Application in the BioSciences* (1993) **9**, No. 3, 729-734.

10. Monge, A.E.: "An Adaptive and Efficient Algorithm for Detecting Approximately Duplicate Database Records," On-line document, URL:http://citeseer.nj.nec.com/monge00adaptive.html, (2000).

11. Monge, A.E., and Elkan, C.P.: "An Efficient Domain-independent Algorithm for Detecting Approximately Duplicate Database Records," SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (1997), Tucson, Arizona.

12. Winkler, W.E.: "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage," *in* Proceedings of the Section on Survey Research Methods, American Statistical Assn., pp. 354-359, (1990).

13. Winkler, W.E.: "Advanced methods for record linkage." Technical report, Statistical Research Division, U.S. Census Bureau, Washington, DC, 1994.

14. Wilcox, L.V., and Magistad, C.C.: "Interpretation and Analyses of Irrigation Waters and the Relative Tolerance of Crop Plants," Bureau of Plant Industry, Soils and Agr. Engineering, Agricultural Research Administration, U. S. Department of Agriculture, Riverside, California, 1943.

15. Stueber, A. M., Saller, A. H., Ishida, H.: "Origin, Migration, and Mixing of brines in the Permian Basin: Geochemical evidence from the eastern central basin platform, Texas." AAPG Bulletin **82**, 9, 1652-1672.

16. Dutton, A. R.: "Hydrogeologic and hydrochemical properties of salt-dissolution zones, Palo Duro Basin, Texas Panhandle preliminary assessment," The University of Texas at Austin, Bureau of Economic Geology Geological Circular No. 87-2, 1987.

LIST OF ACRONYMS AND ABBREVIATIONS

NM WAIDS – New Mexico Water and Infrastructure Data System
GIS – Geographic Information System
TDS – Total Dissolved Solids
NMSEO – New Mexico State Engineer's Office
PWQD – Produced Water Quality Database
ADCS – Automatic Data Capture System
CMT– Corrosion Management Toolkit
NMOCD – New Mexico Oil Conservation Division

APPENDIX 1. COMPILATION OF RESULTS FOR OPERATOR VISITS