UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

LEONARDO HAX DAMIANI

SHPECK - A Geochemical Speciation Modelling Software

Prof. Dr. Carla Maria Dal Sasso Freitas
Advisor

Prof. Dr. Anthony J. Park
Coadvisor

Porto Alegre, September 2015

SUMMARY

LIST OF ABBREVIATIONS AND ACRONYMS 4

LIST OF FIGURES 5

ABSTRACT 6

1 INTRODUCTION 7
1.1 Objectives of this work 8
1.1.1 Emphasize the importance of geochemical speciation modelling 8
1.1.2 Implementation and validation of *SHPECK* 9
1.2 Structure of this work 9
1.3 Summary 10

2 BASIC CONCEPTS 11
2.1 Computer Science Principles 11
2.1.1 Computer Processing and Modelling 11
2.1.2 Software Architecture and Design 12
2.1.3 Software Development 13
2.2 Hydrogeochemistry Principles 15
2.2.1 Introductions to Thermodynamics 15
2.2.2 Hydrochemical processes 21
2.3 Geochemical Modelling 22
2.3.1 Geochemical Speciation Modelling 22
2.3.2 Other Types Of Geochemical Modelling 23
2.4 Summary - TO BE DONE 24

3 COMMERCIAL SOFTWARES REVIEW 25

4 *SHPECK* - SPECIATION MODEL 26

5 CASE STUDY, RESULTS AND COMPARISON 27

6 CONCLUSION 28

REFERENCES 29

LIST OF ABBREVIATIONS AND ACRONYMS

DBH	Debie-Hueckel
GUI	Graphical User Interface
HCI	Human-Computer Interaction
CPU	Central Processing Unit
FLOPS	Floating-Point Operations per Second
GWB	The Geochemist's Workbench
K	Equilibrium Constant
β_i	Stability Constant
IAP	Ion Activity Product
SI	Saturation Index
k_{diss}	Dissolution rate constant
k_0	Pre-exponential (Arrhenius) factor
E_a	Activation Energy
R	Universal Gas Constant
T	Temperature
γ	Activity coefficient
a	Activity
m	Molality
M	Molarity
pH	Power of Hydrogen
I	Ionic Strength
Eh	Redox Potencial
CRUD	Create / Read / Update / Delete
UI	User Interface

LIST OF FIGURES

ABSTRACT

HERE WILL COME THE ABSTRACT

Keywords: Geochemical Modelling, Chemical Equilibrium, Geochemical Speciation, Multiphase System, Software Engineering, Computer Science.

1 INTRODUCTION

Geochemical modelling design the reactions that happen in a geological structure through the usage of chemical properties (either thermodynamics and kinetics) to describe it. The need to understand the Earth's interior (both at high-temperature - magma - and low-temperature - aqueous solutions near the surface) motivates the effort in this area of study with the development of models and simulations. The applications of geochemical models are essential widely in several environmental problems, such as calculating the composition of natural waters, measuring flowing groundwater or surface water and the formation and dissolution of rocks and minerals in geologic formations. A geochemical speciation modelling software is responsible for calculating the distribution of dissolved species between free ions and aqueous complexes and also saturation indexes for different minerals.

Any model requires three major components: specific information describing the point of interest; the equations that drive and solve the model; and the model output. A model is an object represented by a set of mathematical expressions previously thought to represent natural processes and output the results of these calculations - something experimentally verifiable. In this sense, a model is a system capable of prediction which uses observational data as input and produces results of past examination. The thermodynamics and kinetics data used to establish the reactions and mimic the nature are directly responsible for the accuracy and precision of the geochemical model.

In this work, we develop a software that through the stoichiometry formulation calculates the chemical equilibrium of a geochemical system using the approach of imposing mass-balance conditions according to the elements of the system. This process is known as chemical speciation, and the software was baptized as *SHPECK*. It accepts any general combination of elements, species and reactions, allowing the user to create different environments, simulations and, therefore, fully control any aspect and configuration of the model. Also on this work, we show a complete analysis of the available existing solutions and by comparing we made clear the uniqueness of our computer science approach to commonly geochemical modelling problem. With a high-level and object-oriented programming language, we could implement an efficient solution that model the geochemical speciation problem. *SHPECK* contains an interactive and intuitive interface - unique among geochemical speciation software - as well as the support of a built-from-the-ground database structure that handles the management of the whole information that flows inside *SHPECK*. These two contributions are presented as the result of an extensive study about the available software normally in use to perform geochemical speciation simulations.Their flow of information (input and output) are old, complexes and prone to error. Also important to mention that these software fetch the information from flat file databases. Both of these characteristics are responsible for errors, problems and wrong

interpretations.

The principles of chemical equilibrium calculation rely on the law of conservation of mass (also known as the principle of mass conservation), stated by Antoine Lavoisier, and chemical speciation, which was presented on (GARRELS, 1965). The law of conservation of mass establishes that the total mass of an isolated system will remain constant and is independent of any chemical and physical changes taking place within the system. Therefore, the challenge of chemical equilibrium calculations is finding the number of moles that satisfies a system of equilibrium constraints at the moment where forward and reverse reactions rates are the same. These constraints are organized in a form of linear conservation equations, which may be expressed in the form of either linear algebraic atom and charge balance equations or chemical equations (**?**). For the sake of simplicity, in this work we will only deal with chemical equilibrium and not with chemical kinetics calculations since the first one requires only the solution of algebraic equation. It is planned to integrate kinetics reactions in the future.

The system of equations will drive and represent all the interactions between the components of the simulation. Newton's method (also known as Newton-Raphson method) uses the previous guess for the equilibrium calculation in a subsequent step, recursively until find a suitable solution that satisfies the system and the convergences criterias. One must note that the initial guess is generated automatically and used as a seed for the iterations. This method requires the usage of a Jacobian matrix and a residual vector during the algebraic calculations. Geochemical modelling speciation has an important application in processes that occurs in turbidite reservoirs. The process of the water coming from the salt dome contains a high concentration of salts as sodium (Na^+), chlorine (Cl^-) and potassium (K^+). Compactation, cementation, dissolution or recrystalization can be observed inside turbidites when this process happens. These processes might change drastically, for example, the porosity of the rock and, therefore, the storage capacity of oil and gas.

1.1 Objectives of this work

This work has as purpose two main objetives:

1. Emphasize the importance of geochemical speciation modelling and make explicit the need and uniqueness of the new software developed - *SHPECK*. It is a contribution to the geochemical modelling community by the adoption of a structured Computer Science approach.

2. Analysis, comparison, evaluation of the accuracy of the implemented software with the available commercial options as well as demonstrate the advantages of *SHPECK* towards these options.

1.1.1 Emphasize the importance of geochemical speciation modelling

Soils and aquifers are heterogeneous, subsurface systems composed of a large number of components - dissolved salts, minerals, metals, gases, natural organics, microorganisms, animals and plants. The subsurface is one of the most complex systems studied by scientists and engineers today. Because of this, geochemical modelling has gained importance and is being accepted as a useful tool to interpret subsurface geochemical processes. Geochemical speciation is based on thermodynamics concepts and the assumption of chemical equilibrium in geochemical reactions.

1.1.2 Implementation and validation of *SHPECK*

The idea of our own geochemical speciation software has emerged as an application where it would be possible to apply all the physical, chemical aqueous, geochemists and linear algebra concepts and develop a useful tool with intuitive and interactive interface. The most usual approach to the geochemical modelling area is a geochemical modeller that develop a solution to solve his particular problems and generates his code/algorithm - a solution that most of the times is not very reliable and has no scalability. In our case, the computer science team made the necessary efforts to understand and learn all the complex aspects of a geochemical speciation model and from this knowledge, develop a software that will be able to use all the processing power of nowadays computers combined with a solid knowledge in computer architecture, algorithms and software engineering.

1.2 Structure of this work

The rest of this work is structured as follow. In chapter 2, an overview of the basic concepts needed and technical concepts involved in this work. Chapter 3 shows a thoroughly analysis and review of the commercial software available. Chapter 4 deals with the *SHPECK* implementation, precisely and carefully describing the whole system: design options; mathematical treatment and details; implementation and user interface (UI) details; algorithm validation and complexity; architecture and organization of the software as well as the database; data-flow; and iteration control. In chapter 5, it is presented a study case with an interesting and relevant scenario; the results that validates *SHPECK* and a broad comparison between solutions previously addressed in this work. Chapter 6 brings the conclusion of this work. Finally, this work contains an Appendix A, which is a presentation and an analysis of a linear algebra library used for the development of *SHPECK* called *Armadillo C++*. – THIS NEEDS TO BE VERIFIED LATER

1.3 Summary

- The importance of geochemical speciation modelling and simulations: Geochemistry deals with the chemical composition and chemical changes/reactions in the solid Earth and its various components (lithosphere, hydrosphere and atmosphere). Modelling and computer simulation is a valuable tool that can be used to gain understanding of geochemical processes both to interpret laboratory experiments and field data as well as to make predictions of long term behavior.

- Applications and context of geochemical modelling: The motivation of this work is the major issue in simulations of aqueous systems, which is geochemical speciation modelling. Geochemical models are useful to understand several topics, such as the composition of natural waters, the mobility and breakdown of con-taminants flowing groundwater or surface water; the formation and dissolution of rocksand minerals in geologic formations. Several problems that our society has created (and faces now) point outthe need for geochemical modelling: radioactive waste disposal, mining environmentalissues, landfills, and groundwater aquifers analysis. These applications share the needfor geochemical modelling.

- Objectives, differential and contribution of this work: Modelling hydrogeology is sometimes considered not only a science, but also an art. The importance of geochemical modelling and the need for a solid contribution in this area is something are extremely high - proportional to the computing power that had evolved so much in the last couple of decades. *SHPECK* is a watershed that brings together the up-to-date technologies and computing power with the geochemical speciation modelling. In this work we bring a computation approach to push the state-of-art of geochemical speciation modelling by showing that is possible to have an interactive and intuitive interface as well as a structured database consistent with the computational reality of today.

2 BASIC CONCEPTS

At this point, it is important to understand all the different multidisciplinary aspects that are present in the development of a geochemical speciation modelling software. By definition, applying Computer Science to solve problems and create solutions requires to redefine problems outside normal boundaries and generate a new understanding of complex situations by thinking across two or more academic disciplines.

To develop this work, we had to delineate common goals for the different profiles that would take part on it along the way. All of them with a clear view of their roles and with a noiseless communication in any direction. Furthermore, it is vital and benefits crucially the whole work to be able to take advantage of all the different point of views from the diverse professionals profiles participating in this work. All of the mentioned above are fundamental factors to a successful multidisciplinary work.

Therefore, we present a meticulous and detailed review of all the basic concepts necessary to follow the development of this work, both from the computer science side and also from the hydrogeochemistry and geochemical modelling angle.

Along this chapter, we will first address topics of computer science relevant to this work: computer processing and modelling, software architecture and design, and software development. After, we explain the fundamentals hydro-geochemistry principles: an introduction to thermodynamics; and Hydrochemical processes; And to finish we focus on the geochemical modelling with a special section for it. If the reader feels comfortable with these topics, we recommend that you proceed to chapter 3.

2.1 Computer Science Principles

2.1.1 Computer Processing and Modelling

A processor is a small chip that resides in computers and electronic devices. Its job is to receive input, do something with it and provide the appropriate output. Modern processors, which are located inside the *central processing unit* or *CPU*, can handle trillions of calculations per second and even work together to solve really complex instructions. These processors are also known as *cores* and is common to find *dual-core*'s or even *quad-core*'s processors nowadays. Within that *CPU* is an eletronic clock responsible to create series of synchronized electrical pulses. Which are the key to integrate all the computer's components and perform calculations with the data pulled from the memory. In 2013 the supercomputer *NUDT Tianhe-2* performed $33.86 Pflops$. $Pflops$ stands for Peta Floating-Point Operations Per Second and is the regular unit to measure computer performance. This was pointed out here, in order to express clearly the power that we have available nowadays to build applications and model complex systems where count-

less factors influence and drive the process.

The advances in computer processing made possible scientific modelling, which is generate part or feature of the real world to understand, define, quantify, visualize or simulate (HUMPHREYS, 2004). Modelling such systems require a previous knowledge of all the characteristics, the behavior of this domain and what is the goal with this modelling allied with a big *"piece"* of abstraction. Popular models are, for example, conceptual models, operational models, mathematical models, graphical models. The advantages of a model are: help us to communicate; allow us to clarify and test understanding; create credibility and accountability; organize the thoguhts; simplify and solve problems;

These models are based on a series of "orders" that the modeller had clearly expressed earlier. These stack of "orders" are known as algorithms - a series of instructions for how to do something. Instructions that tell the computer how to make decisions and when to do calculations - which are different according to the type of model.

Among the many computer science areas, there is one that studies the complexity of algorithms. As algorithms are programs that perform a series of instructions, complexity analysis allows us to measure how fast a program is when it performs computations. The analysis allows us to explain how an algorithm behaves in the worst case scenario, for instance. This information will come in hand when we analyze the complexity of *SHPECK*'s algorithm on chapter 4.

2.1.2 Software Architecture and Design

Software Architecture is the process of finding a structured solution that achieves all of the technical and operational requirements addressing also attributes as performance, security, value for the user and management. The architecture and design of a software is the *art* of taking into account all of the several factors and take the best path available considering the impact on quality, interface, database structure, performance, maintainability and success of the software. The software architecture is responsible not only by the algorithms and the data structure, but also by the organization, communication, synchronization, functionality and design of the desired elements, scaling and performance. There is no well defined recipe for a good software architecture - it takes time, practice and efforts to start taking the right paths and weighting the options according to the needs. Recognizing paradigms and building relationships among systems can be handful in order to perform succesfully as a software architect. The software architect is responsible for structuring the software with a consolidated and dense foundation - anything other than this implies risk for the application. Studying the scenarios and requirements before designing the application is a must. Poor architecture results in deployment problems, instability, lack of support and the complete failure of the software (sometimes even the complete business). A software architect aims to:

- Catalog all the requirements of the application, as well as the use cases and scenarios;
- Analyze and reduce the risk either for the application and for the business (if there is one involved);
- Be able to adapt the design decisions around the reality - that will most likely change over time;
- Develop a structure where the tradeoffs of all attributes are clear and the impact of any change will be controlled;

During the development of *SHPECK* we adopt the Object-Oriented, also know by its abbreviation *OO* or *OOP*, architectural style. Which is a paradigm based on the division of responsabilities into reusable and self-sufficient objects, each one of them containing the data and the behavior desired to its functionalities and responsabilities. *OO* COMPLETAR AQUI For the purpose of illustration we listed a few other important architectural styles: client/server; domain driven design; service-oriented architecture (SOA);

Before defining the architecture of *SHPECK* we have analyzed points as atribbutes, application type, technologies and deployment options. Only after this section was possible to define which of the design architectures would fit best to our needs - small refinements were done along the way but always building on top of a previous wisely took decision.

2.1.2.1 Software Architecture Principles

The origin of software architecture principples is the need to minimize costs, address properly maintenance requirements and promote the *Seven Basic Principles of Software Engineering* as in (BOEHM, 1983), which are:

1. Manage using a phased life-cycle plan;
2. Perform continuous validation;
3. Maintain disciplined product control;
4. Use modern programming practices;
5. Maintain clear accountability for results;
6. Use better and fewer people;
7. Maintain a commitment to improve the process;

Along these principles, it is mandatory that the software architect or designer to see the large picture of the software that is under his management. The large picture is responsible to make sure that no feature is overlaping (nor duplicating) with another - this will lead to a low coupling and highly cohesive software.

2.1.2.2 Design Principles

Designing a software is composing a structure with different layers responsible for different tasks or properties. This layering must be consistent for any operation and must respect the hierarchy as well as the orientation of this structure. These layers must be connected but never overlapping themselves: duplicating properties/functionalities/responsabilities is a mistake and prone to error. Overlapping layers is the signal of potential inconsistencies and elevated software maintenance costs. Design patterns is also one important term to keep in mind. Establishing a coding style, naming standards and conventions provide a consistent model that will make the software's life longer and more adaptable.

2.1.3 Software Development

Software development is a process that requires extremely careful planning and execution to meet the proposed goals. The proposed goal is a software, but sometimes people forget that to achieve this goal is necessary many hours of computer programming,

documenting, testing, bug fixing and decisions making. Software development may include also research, new development, prototyping, modification, reuse, re-engineering and maintenance. The most interesting and relevant points are going to be discussed in the following sections.

2.1.3.1 Life Cycle of a Software Development Projet

This topic is relevant to make clear that lots of work is done before writting any line of code. We can mention tasks as: requirements definition; functional specification; architecture and design decisions; implementing and testing; software deploy; documentation; and maintenance; There may be additional tasks according to the reality of each software development but the idea of life cycle must be clear be a clear notion in the reader's mind.

2.1.3.2 Software Engineering

The importance of software engineering can be shown with by two definitions that contrast in time: *"The establishment and use of sound engineering principles in order to obtain economically software that is reliable and works efficiently on real machines."*, from (BAUER, 1968); and *"Software engineering is the application of a systematic disciplined, quantifiable approach to the development, operation, and maintenance of software, and the study of these approaches; that is, the application of engineering to software."* by the *IEEE Computer Society's Software Engineering Body of Knowledge* from 2004.

The understandment that overlaps in both quotes is that to achieve a *software*, engineering principles (for example: management issues, documentation, infra-structure, directing teams, scheduling and budgeting) are necessary and will be fundamental to reach that goal.

2.1.3.3 Database

The database is responsible to organize, store and retrieve the data so it can be used efficiently and easily. It is composed by a collection of schemes in a way that it supports processes requiring information which are used by the application's internal operations. They are organized according to their organization approach: relational database; tabular database; distributed database; OO database; and flat file database;

Among several types of database, the software engineering previously done should identify which of them suits better the needs of the software. Details of *SHPECK*'s database are presented in chapter 4.

*2.1.3.4 User Interface (*UI*) and Human-Computer Interaction (*HCI*)*

UI and *HCI* are area from the computer science that are influence by psychology, ergonomics, engineering, graphic design and others. Both areas take into account and are products of how humans interact with computers.

Good *UI* are not user-expensive nor task-expensive, they behave naturally as the extension of the user's needs and desires. The software will easily bring more value to its users if the *HCI* happens in a mutually benefitial way - by reaching the software's goal and not being an embarassment or annoyance for the user. Losses of productivity, efficiency, money and usability are expected consequences from a software that has skipped the preparation parts to the development of *UI* and *HCI*.

UI and *HCI* were extensively studied and analysed along the development of *SH-*

PECK.

2.2 Hydrogeochemistry Principles

2.2.1 Introductions to Thermodynamics

In thermodynamics, equilibrium is a state of dynamic balance where the ratio of the product and the reactant concentrations is constant. There are three general approaches to calculating the composition of a solution at equilibrium (PETRUCCI, 2007).

1. Manipulation of equilibrium constants (K): The final concentrations are achieved by mathematical handling of the equilibrium constants; the idea is to express all the parts in terms of the measured equilibrium constant and initial conditions. Thermodynamics databases contain the value for the equilibrium constants obtained through experiments. Demonstration of this can be found in (KEHEW, 2000). The disadvantages of this method is when using this method for a huge number of reactions it may never converge.

2. Gibbs Energy of the system: At equilibrium, the Gibbs Energy (G) is at a minimum. When the object of the study is a close system - no particles entering nor leaving - the total number of atoms of each element will remain constant, therefore, achieving the minimum free energy. Due to the complexity in demonstrating how this method works, it will be supressed here. An interesting algorithm for equilibrium calculation that uses Gibbs energy is described in (ALLAN, 2015). One of the disadvantages of this method lies in the effect of species which appear only in tiny quantities at equilibrium.

3. Manipulation of mass-balance: The total concentration of species that compose the system is the base for this method, (SMITH, 1980) explains this stoichiometric formulation approach. This method takes into account the stoichiometric approach among the species, which generates a system of non-linear mass-action equations. Mass-balance manipulation is the method chosen for this work, and the details are explained further in this work.

Stoichiometric approaches have two general advantages over non-stoichiometric: in the case of real systems and for multiphase problems - in which singularities can occur in the linear equations (SMITH, 1980). It is important to remind the reader that any of the methods described above are equivalent and can be verified in (ZEGGEREN, 1970)

It is important to mention that any analysis resulting from a water sample must be carefully taken. Any geochemical investigation is useless if the integrity of the water of the solid phase is compromised. Results of interpretation and modelling might be incorrect if the sampling was not done properly. A principal objective is to obtain a water sample with the same chemical composition as those of water in its original environment, for example, an aquifer or a surface water. (DEUTSCH, 1997)

2.2.1.1 Thermodynamic Equilibrium

There are mainly two ways to describe thermodynamic equilibrium reactions: Equilibrium and Kinetic. Both of them formulates a closed system and describe the position of the maximum thermodynamic equilibrium. Equilibrium is the moment where there is no more chemical energy to alter the distribution of mass between reactants and products

in the system. The way to model a reaction depends on its rate: an equilibrium reaction is relatively fast on the mass transport process while the kinetic reaction is slow. Therefore, when applying an equilibrium model to a reaction is assumed that the whole mass transfer happens at the same time when the reactant and product are putted together, and this will configurate an equilibrium situation. If the reaction rate is slow, it requires a kinetic description of the reaction. On this work, it will be addressed only equilibrium reactions. (NORDSTROM, 1986)

Assuming the independent equilibrium reactions:

$$0 \rightleftharpoons \sum_{i=1}^{N} v_{ji}\alpha_i \qquad (j = 1, ..., M) \tag{2.1}$$

where v_{ji} is the stoichiometric coefficient of the *i-th* species in the *j-th* reaction; and M represents the number of reactions and N the number of species, with $M < N$. The sign convention is to assign the stoichiometric coefficient negative for reactants and positive for products. Assuming that all the reactions in the system are in equilibrium, the chemical system must also satisfy the mass-action equations:

$$K_j = \prod_{i=1}^{N} a_i^{v_{ij}} \qquad (j = 1, ..., M) \tag{2.2}$$

where K_j denotes the equilibrium constant of the *j-th* reaction; a denotes the activity of the *i-th* chemical species. The equilibrium constant depends on the temperature of the system; therefore, the equilibrium constant needs to be calculated according to the temperature of the system.

It has been known that the driving force of a chemical reaction is related to the concentration of the constituents that are reaction and the concentrations of the products of the reaction. The law of mass action states that any reaction will proceed to the right (dissolution) or to the right (precipitation) until the mass-action equilibrium is achieved, important to keep in mind that it may take years or even thousands of years for that equilibrium to be achieved and after a disturbance in the system, such as an addition of reactants, removal of products, changes in the temperature or pressure, the system will continue to proceed toward this new equilibrium (if the disturbances are frequent compared to the reaction rate, equilibrium will never be achieved) (FREEZE, 1979). Each of the dissolved species will have one representation of the nonideal behavior of components in the solution, which is called *activity* and is presented in details later on this chapter.

Kinetic descriptions is applicable to any reaction but it is needed necessary to describe reactions that are slow in relation to mass transport. The following reaction has a k_1 and k_2 rates for the forward and reverse reactions, respectively

$$aA + bB \underset{k_1}{\overset{k_2}{=}} dD + eE \tag{2.3}$$

Each ion has a reaction rate related to the stoichiometry and is expressed as

$$-\frac{r_A}{a} = -\frac{r_B}{b} = \frac{r_D}{d} = \frac{r_E}{e} \tag{2.4}$$

where a, b, d and e are stoichiometric coefficients of each one of the ions in the reaction. r_A, r_B, r_D and r_E are reaction rates, and they describe the time rate of change of concentration as function of rate constants and concentration. Each one of them express the rate

of change at the chosen ion as the difference between the rate at which the component is being used in the forward reaction and generated in the reverse reaction and is described as follow

$$r_A = -k_1(A)^{n1}(B)^{n2} + k_2(D)^{m1}(E)^{m2} \tag{2.5}$$

where $n1, n2, m1$ and $m2$ are empirical stoichiometric coefficients. When there are reactions in parallel or series the rate laws are even more complex. The dissolution rate constant ($k_d$iss) of a chemical reaction depends on temperatue. The relation between constant and temperature is given by the *Arrhenius equation*, described as

$$k_diss = A * exp(\frac{-E_a}{R * T}) \tag{2.6}$$

where k_0 is the pre-exponential (Arrhenius) factor, E_a is the activation energy, R is the universal gas constant, and T is the temperature in Kelvin. During the development of *SHPECK*, we will not deal with kinetic reactions.

2.2.1.2 Thermodynamic Equilibrium Constant

The *equilibrium constant* (*K*), also known as *stability constant*, is the value of the reaction quotient when the reaction has reached equilibrium, as stated in equation 2.2. *K* depends only on the temperature and on the ionic strength of the solution. According to known reactions' equilibrium constant value, it is possible to determine the value for at any temperature by a polynomial fitting technique or polynomial regression.

Equilibrium constants are determined by measurements of the relevant concentrations of the species under differing experimental conditions. Concentrations of species can be measured in multiple ways, and the use of these values in modelling requires adjustment to the conditions in the system being modelled. These adjustments, as well as the differences in conditions and different methods for determination, can lead to uncertainty in chemical speciation constants.

Several thermodynamics database are available and really popular nowadays. They include reaction constants, reaction descriptions, solutes, species, enthalpy values, activity coefficient parameters, etc. During the development of *SHPECK* we selected the *Geochemist's Work Bench's* (*GWB*) database - it contains also the values of a 8^th degree polynomial which allows the user to calculate the equilibrium constant to any temperature. Also another source of data used along this work is (PALANDRI, 2004).

In geochemical modelling, the usage of polynomial regression is specifically to calculate the equilibrium constant of the compound at the desired temperature. Polynomial regression is one of several methods of curve fitting, which is a process of constructing a curve that hast the best fit to a series of data points. The polynomial regression is a statistic method that is a form of linear regression in which the relationship between the independent variable *x* and the dependent variable *y* is modelled as an *nth* degree polynomial. In our case, the polynomial regression is necessary in order to acchieve the equilibrium constant for compounds found in the solution system. Polynomial regression is considered to be a special case of multiple linear regression. A polynomial is a function that takes the form

$$f(x) = c_0 + c_1 * x + c_2 * x^2 + ... + c_n * x^n \tag{2.7}$$

where *n* is the degree of the polynomial and *c* is a set of coefficients. Polynomial regression models are usually solved using the method of least squares. Likewise performing polynomial regression with a degree 0 on a set of data returns a single constant

value. It is the same as the mean average of that data. This makes sense because the average is an approximation of all the data points, as shown in figure **??**. The average line mostly follows the path of the data points. Thus the mean average is a form of curve fitting and likely the most basic.

Linear regression is polynomial regression of degree 1, and generally takes the form

$$f(x) = c_0 + c_1 * x \tag{2.8}$$

where c_0 is the y-intercept and c_1 being the slope. Figure **??** shows clearly that the linear regression line running along the data points approximate the data. Mean average and linear regression are the most commom forms of polynomial regression, but not the only.

The next step of polynomial would be the quadratic regression, now the regression becomes non-linear and the data is not restricted to straight lines. With figure **??** is possible to visualize a data with a quadratic regression trend line. Basically, the idea is simple: find a line that best fits the data which is find the coefficients to a polynomial that best fits the data.

Polynomial regression is an overdetermined system of equations that uses least squares as a method of approximating an answer. To understand this, some linear algebra is required.

2.2.1.3 Activity of a solute

Activity (a_i) is *"thermodynamic concentration"* (or informally known as *"effective concentration"*). It is calculated as a product of activity coefficient and concentration (where *i* means the solute involved):

$$a_i = \gamma_i * m_i \tag{2.9}$$

Activity coefficient (γ_i) is a function of ionic strenght (I), which is a measure of the concentration of ions in the solution.

2.2.1.4 Ionic strength

Mathematically the ionic strength of the solution is calculated according to

$$I = 0.5 \sum M_i z_i^2 \tag{2.10}$$

where *M* is the molar concentration of the specie *i* having a charge *z*.When *I* increases, activity coefficients decrease. In very diluted solutions activity coefficient is equals to *1.0* and activity is equal to concentration. The decreasing trend is related to the "cage" of opposite charge particles around ions. There is reversal of the trend in extremely concentrated solutions (brines) because beyond ionic strenght of about $1mol/L$ there is an increase of activity coefficients with increasing ionic strength. This is related to decreasing amount of free water because most of water is already bound around dissolved species. For a matter of explanation, we will calculate the ionic strength of a $CaCl_2$ solution (com-

posed by $0.5mol$ of Ca^{+2} and $1mol$ Cl^{-1}):

$$I = \frac{1}{2}(z_{Ca}^2[Ca^{+2}]) + \frac{1}{2}(z_{Cl}^2[Cl^{-1}]) \tag{2.11}$$

$$I = \frac{1}{2}(2_{Ca}^2[Ca^{+2}] + (-1)_{Cl}^2[Cl^{-1}]) \tag{2.12}$$

$$I = \frac{1}{2}(4 * 0.5 + 1 * 1) \tag{2.13}$$

$$I = 1.5mol/L \tag{2.14}$$

2.2.1.5 Activity Coefficient

There are different methods to calculate γ for ions:

- Debie-Hueckel: They assumed that ions behave like spheres with charges located at their center points. The ions interact with each other by coulombic forces and the result of their analysis is as follows

$$log\gamma_i \quad = \quad -Az_i^2\sqrt{I} \tag{2.15}$$

  where *A* is a constant that is a function of temperature, z_i is the ion charge and *I* is the ionic strength of the solution.

- Davies equations: Is a variation of Debie-Hueckel that can be used when the ionic strength is relatively high. The equation is as follow

$$log\gamma_i \quad = \quad -Az_i^2\left(\frac{\sqrt{I}}{1+\sqrt{I}} - 0.3I\right) \tag{2.16}$$

- B-dot: This model is presented as an activity model based on an equation similar to Davies and parameterized for solutions up to 3 molal ionic strength.

$$log\gamma_i \quad = \quad -\frac{Az_i^2\sqrt{I}}{1+a_iB\sqrt{I}} + \dot{B}I) \tag{2.17}$$

  where *å* is the ion size for each specie and *A, B and* $\dot{B}$ are coefficients that vary with the temperature.

Important to mention that there are other methods available to calculate activity coefficients which are not going to be addressed here. Is important to keep in mind that pure solids have an activity equals to one. Each one of the methods has its advantages and limitations. Debye-Hueckel equations are simple to apply and extensible to include new species in the solution due to the fact that it requires a low number of arguments and specific arguments. Besides, Debye-Hueckel can be applied to the most important temperatures in the field of aqueous geochemist. Important to keep in mind that it works poorly when regarding moderate or high ionic strength. Regarding dissolution and precipitation there is clearly a reaction happening during these processes, which means that some reactions are not in equilibrium.

2.2.1.6 Saturation Index

The saturation index (*SI*) indicates the degree of saturation with respect to a given mineral, in other words, it defines if a reaction will be in equilibrium or not. *SI* is expressed as

$$SI = log(IAP/K) \tag{2.18}$$

when a mineral is in equilibrium with a solution the *SI* is zero, a negative *SI* indicates undersaturation and a positive *SI* supersaturation. Ion Activity Product (*IAP*) is calculated according to

$$IAP = \frac{[C]^c[D]^d}{[A]^a[B]^b} \tag{2.19}$$

where [A], [B], [C] and [D] are activitys of each ion. The interpretation of *IAP* is the following:

- IAP > K : The reaction is progressing from right to left, producing more products. In a ground water solution, the water is supersaturated.
- IAP = K : The reaction is in equilibrium, there is no flow neither to the right nor to the left. In a ground water solution, the water and the mineral are in equilibrium.
- IAP < K : The reaction is progressim from left to right, producing more reactants. In a ground water solution, the water is undersaturated.

With the *SI* approach is possible to predict the reactive mineralogy of the subsurface from the groundwater data without collecting samples of the solid phase and analyzing the mineralogy. If the *SI* for a mineral calculates to be less than zero, the aqueous solutions is undersaturated with respect to that mineral - which corresponds to the fact that the mineral will not precipitate and may dissolve in order to reach equilibrium concentrations. If the *SI* is greater than zero, then the mineral is not reactive and the mineral may precipitate from the aqueous solution (oversaturated). To conclude, when the *SI* is close to zero (it is ok to consider a small range of values to be in equilibrium) , it means that the water is saturated with respect to that mineral (ALLEY, 1993). From the mentioned before, it is possible to state the following:

- SI < 0 : Mineral is undersaturated;
- SI = 0 : Mineral is in equilibrium with the solution;
- SI > 0 : Mineral is oversaturated;

2.2.1.7 Hydrogeochemistry common units

Molarity (*M*), defined as mass in moles in 1 liter of solution and molality (*m*), defined as mass in moles in 1 kilogram of solution. In dilute solution molarity is approximately equal to molality. Concentration in miliequivalents per liter is concentration in milimoles per liter multiplied by charge of an ion.

2.2.2 Hydrochemical processes

2.2.2.1 Acid-Base Reactions

The importance of acid-base reactions is cleary when it is understood its influence on the pH. The pH is a master variable in charge of controlling chemical systems and is described as

$$pH = -log([H^+]) \tag{2.20}$$

where $[H^+]$ is the activity of the hydrogen ion. The interpretation of the values is as follows:

- pH < 7 : acid solution;
- pH = 7 : neutral solution;
- pH > 7 : basic solution;

The acid substance has tendecy to lose protons while a base substance has tendency to gain protons and the interation between acids and bases is called acid-base reactions and is described as

$$Acid_1 + Base_2 \quad = \quad Acid_2 + Base_1 \tag{2.21}$$

The reaction must be understood as that in the forward reaction, the proton lost by $Acid_1$ is gained by $Base_2$ and in the reverse reaction the proton lost by $Acid_2$ is gained by $Base_1$. The strength of an acid or base refers to the proportion of its protons are lost or gained.

2.2.2.2 Complexation and Speciation

A complexation is when an ion that forms by combining simpler cations, anions and sometimes molecules, this process facilitates the transport of potentially toxic substances and form what is called a complex. Due to the importance of this process in contamination problems it has acquired a huge importance in practical and commercial fields. A simple example of complexation is the following

$$Mn^{2+} + Cl^- \quad = \quad MnCl^+ \tag{2.22}$$

Calculation of distribution of metals mong complexes (*speciation*) involves the solution of a series of mass-law transport equations. The mass law equation of the reaction 2.22 is described bellow

$$K_{MnCl^+} \quad = \quad \frac{[MnCl^+]}{[Mn^{2+}][Cl^-]} \tag{2.23}$$

Each one of the complex has a variable associated called stability constant (β_i) and it contains the basic information necessary to determine how the total concentration of a metal in a solution is distributed as a metal ion and the other various complexes possible.

2.2.2.3 Oxidation-Reduction Reactions

Groundwater environemnt's reactions involve transfer of electrons between its components (gaseous, dissolved or solid constituents). As result, there are changes in the oxidation states of the reactants and products. It is important to stress that the oxidation number is a hypothetical charge that an atom would have if the ion or molecule were to dissociate. This state can be different according to the solution. During this work, *redox reactions* (as oxidation-reduction reactions are also known) are not going to be addressed. In order to get deeper understanding on this topic, we refer to (PETRUCCI, 2007)

2.2.2.4 Adsorption and ion exchange

Adsorption systems treat water by adding a substance, such as activated carbon or aluminia, to the water supply. Adsorbents attract contaminants by chemical and physical processes that cause them to *stick* to their surfaces for later disposal. This mechanism is often used to remove contaminants like *arsenic* or *fluoride* (mostly organic contaminants) from water reservoir. Ion exchange work simillarly but it is focused in inorganic contaminants in a particle-free water. Ion exchange is most often used to remove hardness or nitrate (mostly inorganic soluble molecules). During this work, adsorption and ion exchange are not going to be addressed. In order to get deeper understanding on this topic, we refer to (FREEZE, 1979)

2.3 Geochemical Modelling

The geochemical modelling is the design of the geochemical reactions responsible for the migration of dissolved species. Geochemical models can be divided into two groups:

- Geochemical Equilibrium Models: Based on the assumption of thermodynamic equilibrium reached in a relatively short time (no time factor is included in calculation). It takes in consideration only equilibrium reactions.
- Geochemical Kinetic Models: It takes into account also kinetic reactions and includes the time factor. As kinetic data is measured experimentally, there is still a lack of kinetic data available for many geochemical processes.

As mentioned before, this work will focus on the first one - *geochemical equilibrium models*.

Inside geochemical equilibrium models we can mention three divisions: speciation models; inverse models (also called mass balance models); forward models (also called reaction path models); and reactive transport (coupled) models. Regarding the relation to spatial coordinates, geochemical equilibrium models are considered *batch models* - which are basically closed vessels or reactors.

2.3.1 Geochemical Speciation Modelling

Speciation represents modelling based in the equilibrium of the system. A geochemical speciation modelling program calculates the distribution of dissolved species between free ions and aqueous complexes and also saturation indexes for different minerals. Sodium, for example, can be present in water as free ion Na^+, and also in the form of complexes with anions:

$$Na^+{}_{total} = NaCl_{aq} + NaOH + Na^+ \tag{2.24}$$

where Na^{+}_{total} is total sodium concentration from chemical analysis. Na^{+}_{total} is a component (e.g., chemical formula unit used to describe a system) and Na^{+}, $NaCl_{aq}$ and NaOH are species (chemical entities which really exist in the system). Information about the distribution of dissolved species is important, for example, for risk assessment of contamination by metals because toxicity of metails depends on their speciation in solution. Carbonate complexes of metails, for example, are less toxic than their free ions. Saturadio index (SI) is used to determine the direction of geochemical processes. When $SI > 0$ the mineral precipitates from the water and when $SI < 0$, the mineral dissolves in contact with the water, if it is present in solid phase. Field data necessary for input of speciation program are temperature, pH and results of laboratory chemical analysis (results from a sampling of the solution of interest).

Common problems solved using speciation programs are:

- There is a sample with high concentration of dissolved sodium and we need to know the distribution of sodium between Na^{+} and different complexes (for example, $aCl NaCl_{aq}$ or NaOH) because different forms of sodium have different characteristics;

- There are ground water samples that had been in contact with granitic masses and we want to verify the possibility of precipitation of minerals like *Albite* (a plagioclase feldspar mineral whose formula is $NaAlSi_3O_8$).

Note that the details of several available programs are going to be presented and discussed in chapter 3).

The development of our software *SHPECK*, which is a geochemical speciation modelling software will be detailed, presented and thoroughly discussed in chapter 4. Also important to mention that this work is the first work that will completely guide anyone to generate a geochemical speciation modelling software from the ground.

2.3.2 Other Types Of Geochemical Modelling

2.3.2.1 Inverse geochemical modelling

This type of models, also known as mass balance models, are used when chemistry of groundwater and solid phase composition are already known, and reactions that have already happened should be determined. It is used when we have access to 2 hydraulically connected points and composition of solid phase between these points; with these data in hand, it is possible to calculate and produce the reactions that will explain the changes of the water's chemistry. This approach leads to some uncertainties: Stoichiometry of minerals in solid phase is not often well known; solution may be non-unique; and programs can produce several possible models for the same input; An interesting work about inverse geochemical modelling can be verified in (SHARIF, 2007).

2.3.2.2 Forward geochemical modelling

This type of models, also called reaction path models, are used for prediction of water chemistry evolution along a flowline. Initial water chemistry is known and the aim of the program is to predict water chemistry at some point along flow path. This kind of modelling introduce problems regarding kinetic and adsorption data, which are ofter missing and frequently limited.

2.4 Summary - TO BE DONE

- Importance of multidisciplinary problems:
- Analysis of the basic concepts necessary from the geochemical point of view
- Analysis of the basic concepts necessary from the CS point of view
- Summary

3 COMMERCIAL SOFTWARES REVIEW

- Geochemical's software analysis/review: The idea is to do a table comparison of softwares and tech details. Also interesting to mention who maintain (OR NOT) the programs. The software analysis and comparison will be only between:
  - EQ3/6 (GWB): Large database, general screening, info is not exactly about what you want. difficult to use and people don't really understand how it is put together. Their algorithm/input/output is old.
  - PHREEQC: shitty interface, poor input/output, old, defferent way of defining the problem.
  - MINTEQ: Much simpler options on raction treatment, developed by the United States' Environment Protection Agency (EPA).
  - SOLMINEQ: its database is focused on organic materials.
- Academic/literature analysis
- Emphasize again the CONTRIBUTION of my approach
- Summary

4 *SHPECK* - SPECIATION MODEL

- Global vision of a simulator
- Architecture of a simulator
- Mathematical treatment and details
- Technologies being used
- Software engineering
- Program organization
- Complexity of the algorithm
- Data-flow
- User interface description and details
- Database: criterios de comparacao do pq usar um banco de dados estruturado; comparacao de tempo com o arquivo texto; simular pesquisa com os 2 formatos; consulatas elaboradas; concatenacao de elementos com queries; espaco, tempo e expressividade; uso de BD em area nao difundida.
- Summary

5 CASE STUDY, RESULTS AND COMPARISON

- Explanation of the data that will be used to compare the results:
- Results comparison
- Summary

Results from initial studies indicated that the uncertainty for thermodynamic values is much greater around the very dilute range and the more concentrated range, where data for the thermodynamic constants are comparatively sparse.

6 CONCLUSION

Geochemical speciation is critical for understanding the form of chemicals of interest in natural systems. It is crucial in many different aspects of our daily life nowadays: assessing bioavailability, risk to humans and ecosystems. Geochemical speciation models are generaly determined through analytical methods that measure free ions or total concentrations, used in conjunction with thermodynamically based models. As presented on this work, these models rely on the local equilibrium assumption, and on experimentally determined reaction constants. It is clear, though, that if the chemical system does not achieve equilibrium state or if the equilibrium constants are not certain, the geochemical speciation model's prediction show significant uncertainty. This emphasizes the clear need for a proper computation approach while treating with such influential information like a geochemical speciation model. The information flow is tightly connected and any mistake will propagate errors and carry that wrong information until the very end of the model.

- Review of the results emphasizing our approach and advantages
- Geological explanations for the results
- Final discussion about *SHPECK*
- CONTRIBUTION (EXPLICIT AND WELL DESCRIBED)

REFERENCES

R. M. GARRELS AND C. M. CHRIST **Solutions, Minerals, and Equilibria**: Harpers' Geoscience Series. Harper and Row, New York, 1965.

KEHEW, A. **Applied Chemical Hidrogeology**: Prentice Hall, 2000. 368p.

FREEZE, R. A. and CHERRY, J. A. **Groundwater**: Prentice Hall 1979

DOMENICO, P. A. and SCHWARTZ, F. W. **Physical and Chemical Hydrogeology - Second Edition**: John Wiley and Sons Inc.

DALE, NELL B. **Programming and problem solving with C++**: Jones Bartlett Publishers, 2004.

WOLERY, T. J. **Calculation of chemical equilibrium between aqueous solution and minerals:** ***the EQ3/6 software package***: Lawrence Livermore National Laboratory, Livermore CA, U.S.A.

WOLERY, T. J., JACKSON, K. J., BOURCIER, W. L., BRUTON, C. J., VIANI, B. E., KNAUSS, K. G. and DELANY, J. M. **Current status of the EQ3/6 software package for geochemical modeling in Chemical Modeling of Aqueous System**: ACS Symposium series, No 416, p 104-116. American Chemical Society, Washington, DC.

WOLERY, T. J. **EQ3/6, a software package for geochemical modeling of aqueous systems: package overview and installation guide (Version 7.0)**: Lawrence Livermore National Laboratory, Livermore CA, U.S.A.

KHARAKA, Y. K. and BARNES, I. **SOLMNEQ: Solution-Mineral Equilibrium Computations**: NTIS Tech Rept. PB214-899, Springfield, VA, 82p

NORDSTROM, D., MUNOZ, K. **Geochemical Thermodynamics**: Prentice Hall, 200. 477p

PARKHURST, D. L. **User's guide to PHREEQC - A Computer program for speciation, reaction-path, advective-transport, and inverse geochemical calculations**: U.S. Geological Survey Water-Resources Investigations Report 95-4227, 143p

BROWN, D. S. and ALLISON, J. D. **MINTEQA1, an equilibrium metal speciation model: user's manual**: Environmental Research Laboratory, Office of research and development, U.S. Environmental Protection Agency.

ALLISON, J. D., Brown, D. S. and Novo-Gradac, K. J. **MINTEQA2/PRODEFA2, a geochemical assessment model for environmental systems: version 3.0**: Environmental Research Laboratory, Office of research and development, U.S. Environmental Protection Agency.

Qt **Application Framework**: Available at http://www.qt-project.org : Accessed in July 2014.

Arma **Armadillo C++ linear algebra library**: Available at $http$: $//www.arma.sourceforge.net$: Accessed in July 2014.

LEE, L. and GOLDHABER, M **The Geochemist's Workbench Computer Program**:

BETHKE, C. M. **Geochemical Reaction Modeling, concepts and applications**: Oxford University Press, 397p

BETHKE, C. M. **Geochemical and Biogeochemical Reaction Modeling**: Cambridge University Press, 547p

XU, T., SONNENTHAL, E.L., SPYCHER, N., PRUESS, K. **TOUGHREACT User's guide: A Simulation program for non-isothermal multiphase reactive geochemical transport in variably saturated geologic media**: Lawrence Berkeley National Laboratory, Berkeley, California, U.S.

HARBAUGH, A. W., BANTA, E. R., MCDONALD, M. G. **MODFLOW-2000, the U.S. Geological Survey modular ground-water model - User guide to modularization concepts and the Ground-Water Flow Process**: U.S. Geological Survey

NOFZIGER, D. L., WU, J. **CHEMFLO : Interactive software for simulating water and chemical movement in unsaturated soils**

PETRUCCI, RALPH H. **General Chemistry: Principles & Modern Applications**: New Jersey : Prentice-Hall

ZEGGEREN, F. VAN., STOREY, S. H. **The Computation of Chemical Equilibria**: Cambridge University Press

SMITH, W. R. **The Computation of Chemical Equilibria in Complex Systems**: Americam Chemical Society

LEAL, M. M. ALLAN, ET. AL. **A chemical kinetics algorithm for geochemical modelling**: Applied Geochemistry

DEUTSCH, J. WILLIAN, **Groundwater Geochemistry: fundamentals and applications to contamination**: CRC Press

PALANDRI, L. JAMES, KHARAKA, K. YOUSIF, **A compilation of rate parameters of water-mineral interaction kinetics for application to geochemical modeling**: U.S. Geological Survey

SHARIF, M. U. ET. AL. **Inverse geochemical modeling of groundwater evolution with emphasis on arsenic in the Mississippi River Valey alluvial aquifer, Arkansas (USA)**: Journal of Hidrology

ALLEY, WILLIAN M. **Regional Ground-Water Quality**

BOEHM, BARRY W., **Seven Basic Principles of Software Engineering**

HUMPHREYS, PAUL. **Extending Ourselves: Computational Science, Empiricism, and Scientific Method**

BAUER, FRITZ **Software Engineering: A Report on a Conference Sponsored by NATO**: NATO