BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF MECHANICAL ENGINEERING
FAKULTA STROJNÍHO INŽENÝRSTVÍ

INSTITUTE OF MATHEMATICS
ÚSTAV MATEMATIKY

STATISTICAL ANALYSIS OF LASER SPECTROSCOPY MEASUREMENTS
STATISTICKÁ ANALÝZA MĚŘENÍ V LASEROVÉ SPEKTROSKOPII

BACHELOR'S THESIS
BAKALÁŘSKÁ PRÁCE

AUTHOR — **Ľuboš Slávik**
AUTOR PRÁCE

SUPERVISOR — **doc. Mgr. Zuzana Hübnerová, Ph.D.**
VEDOUCÍ PRÁCE

BRNO 2018

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA STROJNÍHO INŽENÝRSTVÍ

Zadání bakalářské práce

Ústav:	Ústav matematiky
Student:	**Ľuboš Slávik**
Studijní program:	Aplikované vědy v inženýrství
Studijní obor:	Matematické inženýrství
Vedoucí práce:	**doc. Mgr. Zuzana Hübnerová, Ph.D.**
Akademický rok:	2017/18

Ředitel ústavu Vám v souladu se zákonem č.111/1998 o vysokých školách a se Studijním a zkušebním řádem VUT v Brně určuje následující téma bakalářské práce:

Statistická analýza měření v laserové spektroskopii

Stručná charakteristika problematiky úkolu:

Metoda spektrometrie laserem indukovaného mikroplazmatu (LIBS) umožňuje prvkovou analýzu vzorků ve všech skupenstvích hmoty. Výsledkem měření metodou LIBS je atomové emisní spektrum. Signál jednotlivých prvků ve spektru je reprezentován ostrými píky (spektrálními čarami). Metodu LIBS lze aplikovat na geologické vzorky. V takovém případě rozhoduje společná přítomnost prvků ve vzorku a tedy i ve spektru o druhu mineralizace studované horniny. Některé druhy mineralizace mohou mít pro těžební průmysl zásadní význam.

Cíle bakalářské práce:

1. Základní popis metody LIBS.
2. Zavedení potřebných statistických pojmů.
3. Analýza reálných dat.

Seznam doporučené literatury:

HAHN, David W. and Nicoló OMENETTO. Laser-Induced Breakdown Spectroscopy (LIBS), Part I: Review of Basic Diagnostics and Plasma—Particle Interactions. Applied Spectroscopy [online]. 2010, 64(12), 335A-336A [cit. 2017-09-07]. DOI: 10.1366/000370210793561691. ISSN 0003-7028.

ANDĚL, Jiří. Základy matematické statistiky. Praha: Matfyzpress, 2011. ISBN 978-80-7378-1620.

Termín odevzdání bakalářské práce je stanoven časovým plánem akademického roku 2017/18

V Brně, dne

L. S.

prof. RNDr. Josef Šlapal, CSc. ředitel ústavu	doc. Ing. Jaroslav Katolický, Ph.D. děkan fakulty

Abstrakt
Cieľom tejto bakalárskej práce je stručne vysvetliť metódu spektroskopie laserom indukovaného mikroplazmatu (LIBS) ako nástroj analýzy prvkového zloženia študovaných vzorkov. Výstupy z prvkovej analýzy (emisné spektrá) sú ďalej spracované pomocou matematických metód regresnej analýzy za účelom nájdenia vzťahov medzi nameranými chemickými prvkami. Tieto dve metódy sú použité na analýzu problému založeného na reálnych dátach, a to nájdenie špecifického vzťahu medzi prítomnosťou uránu a hydrozirkónu v študovanom vzorku uranonosného pieskovca.

Summary
The goal of this bachelor's thesis is to briefly describe Laser-induced breakdown spectroscopy method for analysis of elemental composition of studied samples. Outcomes of LIBS analysis (emission spectra) are further proceeded with mathematical methods of regression analysis. The goal is to find relations between measured chemical elements. These two methods are used for analysing real data based problem, that is to find specific relation between presence of uranium and hydrozirconium in a studied sample uranium-hosted sandstone.

Klíčová slova
metóda laserom indukovaného mikroplazmatu, LIBS, regresná analýza, urán, hydrozirkónium, uranonosný pieskovec, prvková analýza.

Keywords
Laser-Induced Breakdown Spectroscopy, LIBS, regression analysis, uranium, hydrozirconium, sandstone-hosted uranium, element analysis.

Rozšírený abstrakt

Táto práca sa skladá z 3 častí. V prvých dvoch kapitolách pokladáme teoretický základ a pozadie nutné pre výklad v tretej časti, analýze problému na reálnych dátach. V prvej kapitole sa venujeme metóde LIBS, ktorou sme dáta potrebné na štatistickú analýzu získali. V druhej kapitole sa venujeme štatistickým metódam, ktoré sme pri analýze vzorku použili. V tretej kapitole vysvetľujeme pozadie problému, náš postup pri analýze a diskusiu k získaným výsledkom.

Metóda LIBS je spektroskopická analytická metóda používaná na zistenie zloženia materiálu, ktorá pochádza z atómovej emisnej spektroskopie (AES). Je veľmi rýchla, dá sa aplikovať na vzorky akéhokolvek skupenstva, nezostávajú po nej výrazné škody na materiáli a vzorky sa neznehodnocujú. Zjednodušený základný princíp je nasledovný. Máme laser ako zdroj veľkej energie. Týmto laserom mierime na konkrétny bod na našom vzorku. Vyšleme veľkú energiu do vzorku a tým sa vzorok v danom mieste začne zohrievať. Elektróny sa začnú rýchlejšie pohybovať a aby vyrovnali nárast energie, prejdú do vyšších energiových stavov. Atómové väzby sa narúšajú a nastáva ablácia (vyparovanie) hmoty. Vzniká plazma. Žiarenie laseru trvá približne desiatky nanosekúnd. Potom začne plazma chladnúť. Ako plazma chladne, elektróny sa vracajú naspäť do nižších energiových stavov a pritom vyžarujú elektromagnetické žiarenie. Detekčné prístroje toto žiarenie zachytávajú v podobe tzv. emisného spektra. Emisné spektrum je tvorené v našom prípade vlnovými dĺžkami v rozmedzí 198.7959 - 1016.708 nm. Na každej vlnovej dĺžke sa zachytáva bezrozmerná intenzita žiarenia. Z kvantovej teórie vyplýva, že každému chemickému prvku odpovedá tzv. charakteristické žiarenie vznikajúce pri prestupe elektrónov z vyšších energiových stavov do nižších energiových stavov. Preto spätne z emisného spektra dokážeme určiť, žiarenie akých prvkov sme zachytili. Teda ak sme na danej vlnovej dĺžke zachytili pomocou detektoru intenzitu väčšiu ako pevne zvolená hranica, môžeme prehlásiť, že v študovanom vzorku sa odpovedajúci chemický prvok nachádza. Polohy vlnových dĺžok odpovedajúcich jednotlivým prvkom sú organizované do tzv. spektroskopických tabuliek. Tú, ktorú v práci používame my, spravuje Americký úrad pre štandardizáciu a testovanie (NIST).

Regresná analýza je štatistická metóda ktorá popisuje príčinno-následné vzťahy medzi náhodnými veličinami. Teda ak existuje príčinno-následný vzťah medzi dvoma veličinami, zmena hodnôt jednej veličiny vyvolá zmenu hodnôt druhej veličiny. Na základe takéhoto vzťahu dokážeme pomocou regresnej analýzy vytvárať predikcie. Pre to musíme zostaviť model v ktorom máme na jednej strane predikovanú náhodnú premennú ktorú vysvetľujeme, a na druhej strane vysvetľujúce náhodné premenné, ktorými vysvetľujeme vysvetľovanú náhodnú premennú. Takýto model sa dá zostaviť mnohými spôsobmi, a preto na nájdenie toho najvystihujúcejšieho modelu používame testy štatistických hypotéz. Rovnako tak je potrebné testovať samotný model na splňovanie predpokladov regresnej analýzy, ako sú nezávislosť reziduí, ich normalita, či konštantný rozptyl.

Štatistickú analýzu sme robili na dátach získaných pomocou LIBS metódy z reálneho vzorku. Naším vzorkom je kúsok uranonosného pieskovca získaného zo Stredočeského kraja. Analýza tohto vzorku je dôležitá pre ťažbu uránu ako dôležitý zdroj energie pre jadrové elektrárne. Celková problematika sa zaoberá tým, akým spôsobom sa urán vyskytuje v uranonosnej rude, resp. s akými prvkami sa viaže. Analýzou týchto vzťahov dokážeme lepšie lokalizovať miesta vhodné na ťažbu uránu. My sme sa v práci zamerali na vzťah uránu a fázy zvanej "hydrozirkón" – $ZrSiO_4 \cdot nH_2O$ – spojenie dvoch chemických prvkov – kremíka a zirkónu. Zisťovali sme, či prítomnosť uránu závisí jednotlivo na

zirkóne či kremíku, alebo závisí práve na ich spojení vo fáze hydrozirkón. Analýzu sme robili na dátovom súbore vzorku zloženého z 22 500 emisných spektier, ktoré vznikli meraniami na štvorcovej mriežke o rozmeroch 150 × 150 bodov. Postupovali sme naslednovne.

Na začiatku sme zostavili počiatočný úplny model zložený z vysvetľovanej premennej intenzity uránu, a s vysvetľujúcimi premennými všetkými samostatnými prvkami (7 chemických prvkov) až do tretej mocniny, spolu s interakciami medzi každými dvoma prvkami. Vznikol model s 88 regresnými parametrami. Testovanie modelu ukázalo, že väčšina parametrov je nevýznamná, taktiež grafy reziduí nespĺňali predpoklady. Z tohto modelu sme preto postupne vyradili nevýznamné premenné tak, aby sme dostali iba model s významnými premennými, ktorý by mal dostatočne vystihovať realitu.

To sme urobili takým spôsobom, že v každom kroku tohto postupu sme z modelu odstránili jednu premennú, ktorá odpovedala najvyššej p-hodnote t-testu nulovosti príslušného regresného parametru. Po každom odobraní premennej sme nový submodel ešte otestovali na rovnosť rozptylov s pôvodným základným modelom pomocou F-testu. Vďaka tomu submodel nestrácal pôvodnú kvalitu a na druhej strane získaval na štatistickej jednoduchosti. Tento postup sme opakovali až dovtedy, kým všetky regresné parametre boli štatisticky významné na hladine významnosti 0.05 (p hodnota bola menšia ako 0.05) alebo výsledok F-testu bol, že zamietame rovnosť rozptylov (p hodnota bola menšia ako 0.05). Keď sme dosiahli jednej z týchto dvoch hraníc, submodel sme považovali za najlepší submodel s daným pôvodným základným modelom. Taktiež sme priebežne sledovali, či sa mení rozptyl a normalita reziduí.

Rovnaký princíp sme potom aplikovali pri modeloch s transformáciami vysvetľovanej premennej alebo vysvetľujúcich premenných. Vyskúšali sme transformovať predovšetkým vysvetľovanú premennú, a to odmocninou, mocninou, prirodzeným logaritmom a lomenou funkciou. Taktiež sme skúšali rôzne spracvoať celý dátový súbor. Keďže spôsob získavania dát z metódy LIBS nie je jednoznačne stanovený, existuje viacero prístupov ako dáta prvotne spracovať. Okrem pôvodných "surových" dát sme vyskúšali aj celkovú štandardizáciu a lokálnu štandardizáciu dát. Takto upravené dáta sme rovnako tak otestovali pre rôzne modely.

Najlepší výsledok, ktorého sme boli schopní dosiahnuť, bol pre neupravené vstupné dáta, v modeli s odmocninou vysvetľovanej premennej. Tento model sme boli schopní po obstránení nevýznamných regresných parametrov zúžiť na 52 regresných parametrov, pri nezamietnutí hypotézy o rovnosti rozptylu s rozptylom základného modelu. Submodel obsahoval všetkých 7 pôvodných chemických prvkov, 22 interakcií, a 48 regresných parametrov je v modeli štatisticky významných. Koeficient determinácie, teda schopnosť modelu vysvetliť variabilitu vysvetľovanej premennej je na úrovni 98.76%. Podľa grafu sú reziduá homoskedasticitné, teda majú rovnaký rozptyl, a taktiež ich rozdelenie je veľmi podobné normálnemu rozdeleniu. Takýto výsledok, predovšetkým grafický, sme nedokázali dosiahnuť pri žiadnom inom modeli.

Následne sme tento model analyzovali vzhľadom k pôvodnému problému, a to akú rolu v ňom zohrávajú zirkón a kremík, vrátane ich vzájomnej pôsobnosti. V najvhodnejšom submodeli vystupuje kremík v prvej a druhej mocnine, zirkón v prvej, druhej a tretej mocnine, a zároveň aj ich spoločná interakcia. Okrem druhej mocniny kremíka sú všetky tieto prvky štatisticky významné. Pri druhej mocnine kremíka túto skutočnosť neberieme ako smerodajnú, a to hlavne preto, lebo v modeloch komplexných ako tento sa môže stávať, že vysvetľujúce premenné ako samostatné parametre sa môžu javiť ako nevýznamné, no smerodajné je, ako sa správajú vo svojich interakciách. Pri kremíku

môžeme pozorovať 4 štatisticky významné interakcie, preto usudzujeme, že kremík je v modeli významný. Štatisticky nám teda tento submodel nevyvrátil, že by sme mohli pri vysvetľovaní uránu zanedbať samostane kremík, zirkón, alebo ich interakciu. Fáza hydrozirkónu v našom vzorku sa teda javí ako významná, a preto prítomnosť uránu nezávisí od zložiek hydrozirkónu samostatne, ale s fázou ako celkom.

Takto získané výsledky sme sa snažili ešte podporiť zostavením ekvivalentného glm modelu. Aplikovali sme rovnaký princíp odstraňovania premenných spolu s F-testom. Získaný submodel sme porovnali s naším najlepším lineárnym submodelom. 17 z 22 interakcií sa nachádzalo ako v lm modeli, tak aj v glm modeli, čo považujeme za podporu naších úvah. Následne sme porovnali predikcie týchto dvoch modelov. Glm model predikuje vyššie hodnoty uránu ako lm model pri hodnotách intenzít do hranice jedna, zatiaľ čo lm model predikuje vyššie hodnoty ako glm model nad hranicou intenzít jedna. Všetky získané výsledky sú v práci hlbšie okomentované aj spolu s grafickými ukážkami.

SLÁVIK, Ľ. *Statistical analysis of laser spectroscopy measurements.* Brno: Vysoké učení technické v Brně, Fakulta strojního inženýrství, 2018. 53 s. Vedoucí doc. Mgr. Zuzana Hübnerová, Ph.D.

I declare that I have written the Bachelor's thesis *Statistical analysis of laser spectroscopy measurements* on my own and under the guidance of my Bachelor's thesis supervisor doc. Mgr. Zuzana Hübnerová, Ph.D. and professional consultant Ing. Jakub Klus, Ph.D., and using the sources listed in references.

Ľuboš Slávik

I would like to express huge gratitude to two most important people for completing this bachelor's thesis. To my supervisor doc. Mgr. Zuzana Hübnerová, Ph.D. for her unbelievably positive attitude and proactive approach to this work, beyond thesis itself, and especially toward me. Without your guidance and help I wouldn't be able to finish this thesis and make such a great personal progress. To my professional consultant Ing. Jakub Klus, Ph.D. for his eternal patience with explaining all the physics behind LIBS, for his professional scientific viewpoint humanly delivered, and for his precious text feedback that made this thesis look at least little bit scientific.

Ľuboš Slávik

Contents

1 Introduction

Analysing composition of things around us is nowadays absolutely common element of modern lifestyle. In stores we're checking what products we buy consist of, so we know what ingredients we are putting inside our body. When a doctor prescribe us a medicine, we ask "What's inside of the pill?" so we can search on the internet what is the active pharmaceutical ingredient and what effect does it have, especially what side effects might the medicine have. Lastly, when you visit a bookstore, you don't just look at the book cover, say "Yeah, such a nice book, I'm okay now." and leave. You want to open the book, read through it and know what is it about. That's a human nature, to go inside things, to understand what's behind what we can see, smell or touch.

Here in AtomTrace at Brno University of Technology, where this bachelor thesis was created, we're trying to look deep inside the materials and analyse what chemical elements our samples do consist of. We're developing measurement tool that would be able to easily analyse chemical composition of materials of any sort in various environments under all conditions. As a measurement method they're using Laser-Induced Breakdown Spectroscopy.

Laser-Induced Breakdown Spectroscopy (LIBS) is quite a new method for element analysis. This method has many advantages such as speed, demands at sample preparations, applicability to all states of matter. The outcome of LIBS measurement is LIBS spectrum. This spectrum contains a huge amount of data. That's where statistics takes its place.

Regression analysis is one of the most common statistical methods used in wide range of professions such as economists, analytists, data scientists, social workers, or psychologists. Its capabilities to capture relationship between things and describe the cause-effect relation is in this era of data analysis priceless. Also properties of prediction abilities have huge potential in avoiding accidents, for example in medicine.

In this thesis we're analysing presence of Uranium in sandstone-hosted uranium deposit. In addition we're searching for particular relationship between Uranium and so called "hydrozirconium". Exploring its behaviour and principles of occurrence can help in mining of Uranium, which is important in creating energy needed for our everyday lives.

2 Laser-Induced Breakdown Spectroscopy

Laser-Induced Breakdown Spectroscopy (LIBS) is new, developing field in chemical analysis. Method has roots in Atomic Emission Spectroscopy (AES) and recently, it's becoming popular. Mostly regarding its capabilities to analyse samples in all states of matter, with high repetition rate, no physical contact, and without almost any destruction of the sample (leaves just little craters almost not visible to human eye). Same counts for number of publications that is rising from year to year. For its potential this technique had been called "a future star" [4] by Dr. James Winefordner, world known spectroscopist.

This chapter explains what exactly LIBS system is, how does it work, what components does it consist of, what can be regulated by a person and finally what is the outcome of the measurement. This measurement is used for our data analysis. Sources used for this chapter are [11], [16], and [17].

2.1 Basic concept

To understand the basics of how LIBS works, there's no need to have deep knowledge in chemistry, physics, or technology in general. It's just necessary to accept some logical assumptions.

Imagine a laser. Tool that is able to concentrate energy (have high density of photons) and out of this energy produces a beam of intense radiation (light). The radiation is so intense it can burn person's retina in a second. Very similar to if a person would look directly to the sun for about 30 seconds. Furthermore, this light is also strong enough to "burn" any material. "Burn" means it delivers so much energy to the targeted point on the sample, that particles of the material start to rapidly move and change their "state". Now a physical processes – ablation, plasma creation, excitation – take place. For now just assume that the plasma is created on the surface of the sample and when it cools down, it also emits electromagnetic radiation. This radiation is collected and measured by detectors. Detected signal is called LIBS (or emission) spectrum which is represented by measured intensity dependent on wavelength usually showed in a plot. On the x-axis there is wavelength (between 198.7959 - 1016.708 nm) of the radiation and on the y-axis it's intensity (non-negative dimensionless quantity) of the radiation. According to Atomic Emission Spectroscopy every element, after this process of ablation and when plasma starts to cool down, emits energy at certain wavelength. So when we look at emission spectrum, we find many peaks[1] that represent certain chemical elements. By analyzing these peaks we conclude which chemical elements occur in the targeted place of the sample. From the moment when laser produces a beam till the moment of detecting emission spectrum, whole process takes about tenths of microseconds. On the other hand, one light pulse can analyze only a very small targeted volume of the sample, which means the measurement has to be repeated on more places and the outcomes (LIBS spectra) have to be processed, most preferably statistically evaluated.

[1] Since there is only limited amount of elements in the sample, emission spectrum theoretically looks like a constant line with some points of high intensities – peaks – that lie above this line and are significantly higher than their close and overall background.

2.2 Atomic Emission Spectroscopy (AES)

LIBS method's roots reaches to older analytical method developed in the end of 19th century. First to experiment with this method were Kirchhoff and Bunsen [13]. AES is an analytical method used for determining composition of samples in all states of matter. The main process consists of destructing chemical bonds of the sample (atomization and excitation)[2]. For example, Kirchhoff and Bunsen used just a simple flame. Other tools for destructing chemical bonds might be electrodes, arch or spark discharge, hollow cathode lamps [19], inductively coupled plasma [17] or as in LIBS a laser. When the energy is delivered to the interaction spot on targeted sample, temperature can rise up to around 30 000 K, chemical bonds are destroyed and atoms get to the higher energy levels[3] to compensate this gain in energy as shown in Fig. 2.1.

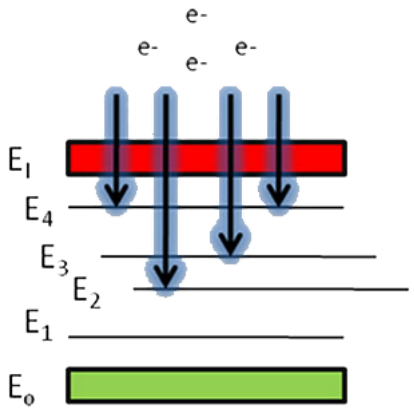


Figure 2.1: Emission of electrons in atom. e are electrons that are moving to higher energy states $E_1 - E_4$ [8].

After process finishes and plasma starts to cool down (electrons get back to lower energy levels), atoms emit radiation which is detected. AES shows us, that this radiation is unique for every element. Therefore we can assign this radiation to its origin – chemical element. Information about which wavelength belongs to which chemical element can be found at NIST database [2]. In specific cases when the instrumentation is properly calibrated and environment is optimal, intensity can also provide the quantification of the elements [5].

2.3 Laser light

Light Amplification by Stimulated Emission of Radiation (Laser) is the basic component for executing LIBS. Laser appears as energetic source that stimulates radiation and creates great intensity of energy on the sample surface. There are many parameters that can be defined upon which different results appear on the sample. We'll state the most important ones here.

Light power and *pulse energy* are highly adherent terms. They express how strong the pulse is, how much energy it provides:

[2]Process of releasing atoms from chemical bonds and delivering energy to them to get to the higher energy states.

[3]Electrons get to the higher shells (further from the nucleus) as the result of receiving energy.

$$P = \iiint I_{pulse}(r, \lambda, \varphi) d\varphi d\lambda dr, \tag{2.1}$$

$$E = \iiiint I_{pulse}(r, \lambda, \varphi, t) d\varphi d\lambda dr dt, \tag{2.2}$$

where P is light power [J], E pulse energy [J], I is intensity of the pulse, r is space vector, λ is wavelength, φ is polarization angle and t is time. There is also a relation between power, intensity and energy:

$$E = \int_{pulse} P(t) dt, \tag{2.3}$$

where light power is integrated by the duration of laser pulse.

The way laser delivers energy is by pulsing and focusing energy. Therefore the pulse duration and also pulse shape are essential for the proper description. Most common pulse profiles are shown in Fig. 2.2.

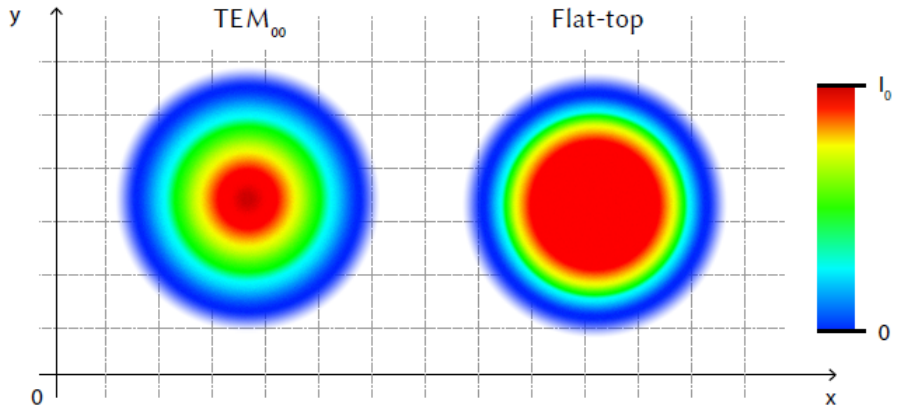


Figure 2.2: Crosswise profile of laser pulse in Gaussian shape (TEM_{00}) and flat-top. I_0 is maximum pulse intensity.

2.4 Plasma

When a laser light impacts the sample, regarding laser light's energy, surface of the sample is heated to around 30 000 K and some of the material (nanograms) is ablated[4] from the sample and ablation crater is created, which diameter is being beneath 100 µm. Atomization and ionization happens (as explained in 2.2). Thanks to delivered energy atoms are excited and free electrons start to crash into each other. This process spreads and laser generated plasma is created. Plasma is a ionized gas with increased electron density and heat. It typically behaves uniformly and appear almost neutral. After the laser pulse stops (takes around 10 ns), plasma starts to cool down (takes around 30 µs) and leave crater and redundant material behind. Whole process cover up complex particle interactions and many parameters are to be included. Plasma cycle is shown in Fig.2.3.

[4]Ablation is a process of mass removal (vaporization) from the material by absorbing laser energy. Direct consequence is plasma creation.

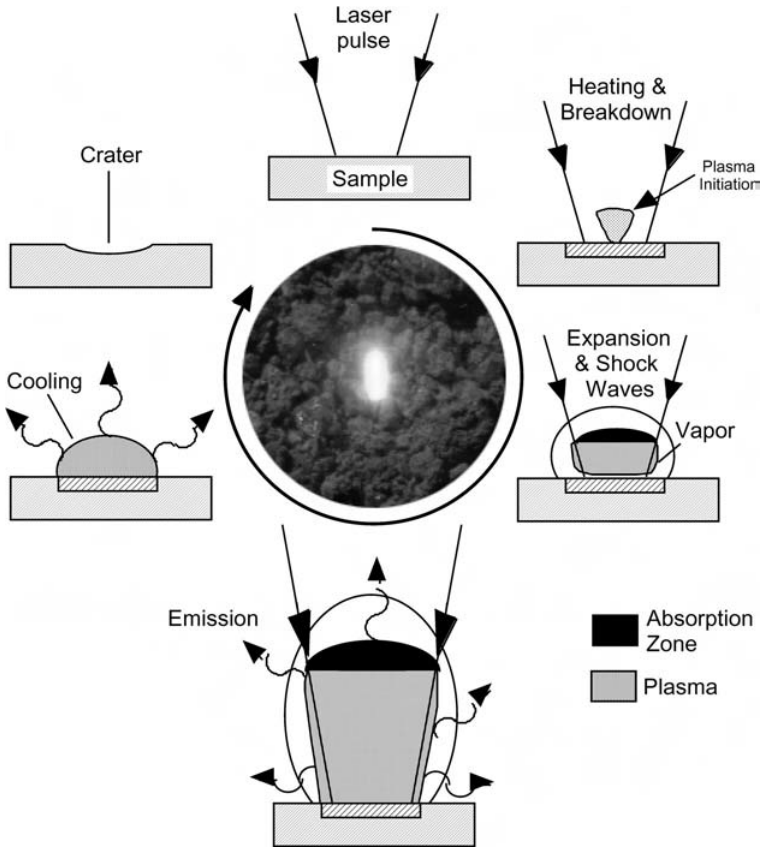


Figure 2.3: Cycle of plasma creation and deconstruction [16].

Here are the most important parameters observed for plasma.

Ablation efficiency is defined as the amount of mass removed per unit energy delivered to the sample. There are more ways of formulating the ablation efficiency, we'll state here only the definition by Vogel and Venugopalan [20], who did the experiments on biological tissue. Other ways of understanding ablation efficiency are as volume of matter ablated to the laser pulse energy or as the ratio of the crater depth to the laser fluence[5]:

$$\eta_{abl} = \frac{\rho\delta}{\Phi_0}, \tag{2.4}$$

where ρ is density of the tissue, δ is etch depth, Φ_0 is called radiant exposure and is equivalent for laser fluence.

Mass removal rate is an expression refering to laser ablation of foils and laser micro-analysis:

$$\dot{m} = \frac{\rho_0 d_T}{\tau_L}, \tag{2.5}$$

where ρ_0 is target density, d_t is foil thickness and τ_L is pulse width.

2.5 Emission spectrum

When we deliver energy with a laser to the sample, plasma starts to evolve. First it heats up, atoms are emitted, and as plasma is cooling down, ablated particles emit characteristic

[5]Fluence is referred to be a time-integrated intensity or time-integrated irradiance.

electromagnetic radiation. Signal detected is time dependent, therefore in course of time atomic lines detected appears differently. After 300 ns first atomic lines are formed. Best time to observe detected radiation is highly depending on many other factors and parameters, such as instrumentation setting used and experimental conditions. In most cases it is between 1.5 - 12 µs. How the signal evolves in time is shown in Fig.2.4.

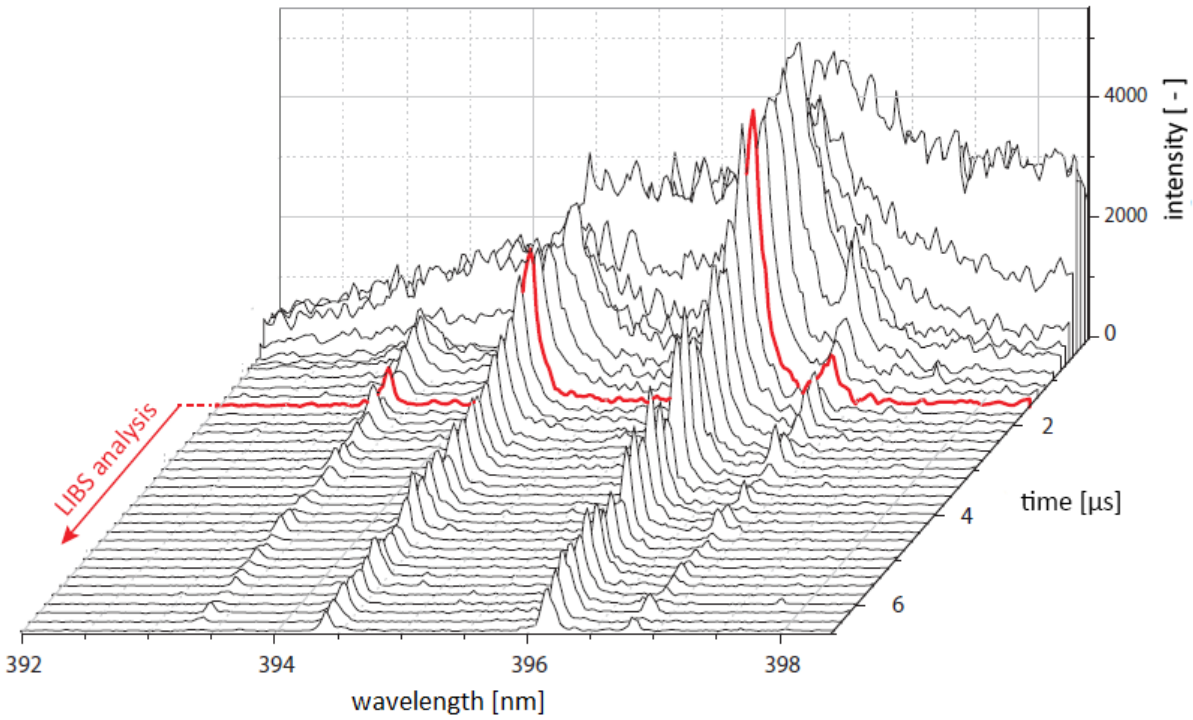


Figure 2.4: Spectrum development over time with step of 100 ns and exposition time 500 ns. Red line refers to best time for executing LIBS measurement [17].

When we choose suitable atomic emission line, we get the data for analysis. These data consist of wavelength[6] and its intensities. If there is a certain element in our sample, that should reflect to high intensity at characteristic wavelength an element emits energy. Fig. 2.5 shows us how these high intensities – elemental lines – appear in the spectrum.

How to determine which lines belong to which elements had been described in 2.2. Since the measurement is not ideal (in which case the peak would appear as a single point with high intensity on the certain wavelength), we have to deal with interpreting and handling the spectrum outcome, i.e. excluding the noise, taking averages of intensities in time period. For detecting elements needed for our analysis, in our measurement we took 0.1 surrounding at λ_e of elements we wanted to describe, i.e. $< \lambda_e - 0.1, \lambda_e + 0.1 >$ and summed measured intensities "in" this neighbourhood. Extracted element's intensities are inputs for the following statistical analysis.

[6]In our measurements we gathered the data for wavelengths between 198.7959 - 1016.708 nm.

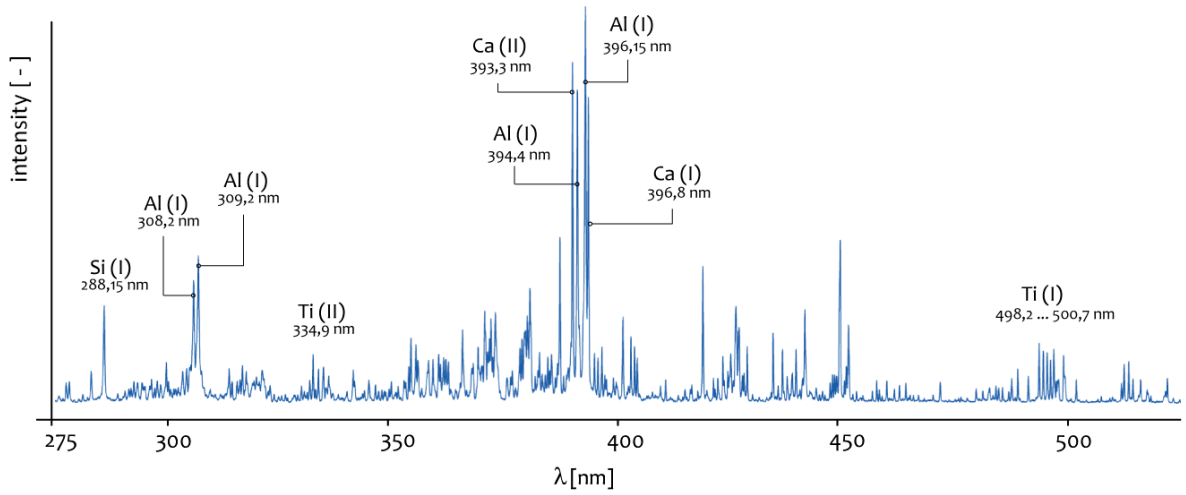


Figure 2.5: LIBS spectrum of ceramic sample with significant elemental lines [17].

2.6 Laboratory equipment and experimental parameters

The LIBS laboratory equipment at Brno University of Technology used for obtaining the experimental data processed in this thesis is listed below.

- Primary laser – High energy **Nd:YAG laser LQ 529A** (Solar LS, BY), operates on its second harmonic (532 nm, 12 ns pulse duration), was introduced into the LIBS chamber by a series of mirrors and then focused by a 25 mm focal length glass triplet (Sill Optics, DE) collinearly with the sample surface normal.
- Secondary laser – **Brilliant B** (Quantel, FR) operates at the fundamental wavelength (1064 nm, 8 ns pulse duration), was introduced into the chamber perpendicularly with respect to the first laser pulse axis using mirrors and then focused into the emerging plasma with 40 mm focal length lens.
- Reflective optics – **CC52** (Andor, UK) collects radiation of luminous laser-induced plasma and also via optical fibre (Ø40 m, Thorlabs, USA).
- Camera – ICCD camera **iStar 734i** (Andor, UK; 1024 x 1024 pixels, effective pixel size 19.5 ×19.5 μm), spectrally resolving radiation.
- Spectrometer – **Mechelle 5000** (Andor, UK; 200 - 975 nm, F/7, 6000 /Δ).
- Pulse generator - **DG535** (Stanford Research System, US), control the gate delay and gate width together with special electronics developed in the laboratory of Brno University of Technology.

Experimental parameters used were:

– Ablation laser energy: 30 mJ

– Secondary laser energy: 80 mJ

- Spot size: 50 µm
- Interpulse delay: 0.5 µs
- Gate delay: 1.5 µs
- Gate width: 20 µs
- Spatial resolution: 100 µs

Visualisation of the apparatus is shown in Fig. 2.6 below.

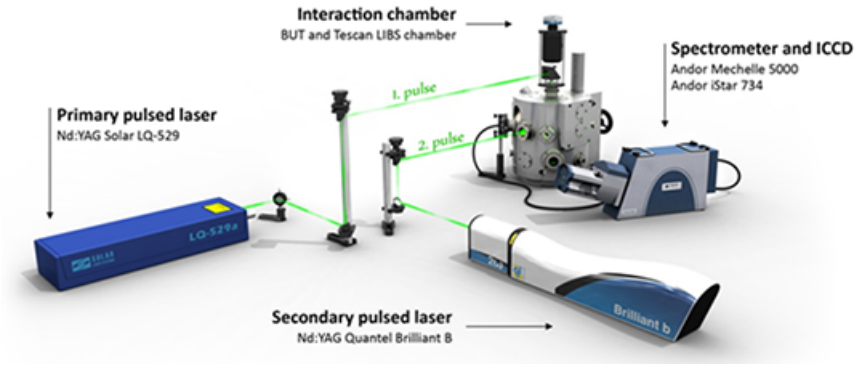


Figure 2.6: Main components of LIBS setup at Brno University of Technology [3].

3 Regression analysis

Most known and very useful tool in statistical analysis is regression analysis. Regression was introduced by Francis Galton [9], who did a research of the relationship between a height of the parents and their children. Based on observations from the real world, the purpose of the regression analysis is to help us identify relationships between things, but mostly to determine an existence of cause and effect between those things. The world is connected and almost all the things come together with something else. If you stay in the sun for too long, you will get burnt. If you shout too much, your voice will hurt next day. Very trivial. But what about, if you listen to music often at certain high volume, what probability you have of becoming deaf? If a patient comes to a hospital with breathing difficulties, high blood pressure, dizziness, is he going to have a heart attack? What is the probability? Do these symptoms even initiate a heart attack? Regression analysis can help us understand how things are connected, and if the relationship seems to be explaining the reality, it is even able to predict future behaviour. It is important to realise that regression analysis serves the purpose of exploring the world, natural laws and connections between things, how they influence each other. In this chapter, I will describe some basic principles, assumptions, and explanations of regression analysis based on [1], [18], [7] and [15].

3.1 Regression model

Creating a regression model means establishing an assumed relationship between random variables called *predictor variables* and *response variables*. Predictor variables are variables, that can be set to a certain value (in experiments) or they are observations of the real world and we can not control them. If there is a change in predictor variable, this change project also into the change of response variable. This is called causality. For the purpose of different terms being used in publications, we understand equality in terms as follows.

$$\begin{aligned}\text{Predictor variables} &= \text{input variables} = \text{inputs}\\ &= \text{X-variables}\\ &= \text{regressors}\\ &= \text{explanatory variables}\\ &= \text{independent variables}\end{aligned}$$

$$\begin{aligned}\text{Response variables} &= \text{output variables} = \text{outputs}\\ &= \text{Y-variables}\\ &= \text{dependent variables}\end{aligned}$$

Therefore the regression model is usually expressed in a form

$$\text{Response variable} = \text{Model function} + \text{Random error},$$

where the model function is a linear combination of X-variables and function depends on unknown parameters.

Let $Y_1,\ldots,Y_n$ be n observed response values and *design matrix* $\boldsymbol{X} = (x_{ij})$ of type $n \times k$, where $k < n$, k is number of predictors (columns), n is number of observations, and $\boldsymbol{X}$ has full column rank (this assumption holds for whole chapter). We assume that there is a relationship between vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ and $\boldsymbol{X}$ defined by equation

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \tag{3.1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ is vector of unknown parameters and $\boldsymbol{e} = (e_1, \ldots, e_n)'$ is random vector called *random error* and meets conditions $E(\boldsymbol{e}) = \boldsymbol{0}$, $var\boldsymbol{e} = \sigma^2\boldsymbol{I}$ and $\sigma > 0$ is also unknown parameter. This model is called *linear regression model*.

Usually, the first column of design matrix $\boldsymbol{X}$ is made of ones. Then we write $\boldsymbol{X_p} = \boldsymbol{(X_0, X_k)} = \boldsymbol{(1, X_k)}$, where $\boldsymbol{X_k}$ is matrix $n \times k$, $\boldsymbol{X_0}$ is a column vector of length n made of ones, and $p = k + 1$. Thus equation (3.1) can be written as

$$\boldsymbol{Y} = \boldsymbol{\beta_0} + \boldsymbol{X_k}\boldsymbol{\beta} + \boldsymbol{e} \tag{3.2}$$

or

$$\boldsymbol{Y} = \boldsymbol{X_p}\boldsymbol{\beta} + \boldsymbol{e}. \tag{3.3}$$

3.1.1 Least Square Estimation

The basic method used for estimating the best line fitted to observed values is least square method. Assume that $\boldsymbol{Y}$ follows linear regression model from 3.1. Equation (3.1) can be written as:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ik} + \epsilon_i \qquad i = 1, 2, ..., n, \tag{3.4}$$

where x_{ij} is ith value of jth variable (column) of $\boldsymbol{X}$ $(j = 1, 2, ..., k)$ and it is known value, Y_1,Y_2,...,Y_n are n observed responses of Y, and β_r $(r = 0, 1, ..., p)$ are unknown regression parameters to be estimated by least square method. Therefore regression model (3.1) can be expressed by n equations as:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} x_{10} & x_{11} & x_{12} & \cdots & x_{1,p} \\ x_{20} & x_{21} & x_{22} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n0} & x_{n1} & x_{n2} & \cdots & x_{n,p} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}, \tag{3.5}$$

where $x_{10} = x_{20} = \cdots = x_{n0} = 1$ is the first column of design matrix and represent intercept of the regression model. If β_0 is the only parameter that Y variable depends on, we claim that response variable is a random sample, thus it does not depend on any variables. For estimation of parameters by least squares, we demand $\boldsymbol{(Y - X\beta)'(Y - X\beta)}$ to be minimal as a function of $\boldsymbol{\beta}$. If design matrix $\boldsymbol{X}$ has full rank, this solution is unique. Otherwise, estimations of $\boldsymbol{\beta}$ are not uniquely defined. Property of best estimation is, that it has the smallest variance from all possible unbiased estimations of linear type.

Theorem 3.1 *Best least square estimation of* $\boldsymbol{\beta}$ *is* $\boldsymbol{b} = \boldsymbol{(X'X)^{-1}X'Y}$.

Proof. See [1] page 81. □

The system of equations $\boldsymbol{X'Xb} = \boldsymbol{X'Y}$ is called *normal equations*. Solved for $\boldsymbol{b}$ we obtain the vector of predictions $\hat{\boldsymbol{Y}} = \boldsymbol{Xb} = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$ which is considered to be the best approximation of vector $\boldsymbol{Y}$, that is possible to create out of a linear combination of columns $\boldsymbol{X}$. A straight line fitted to a data is in Fig. 3.1.

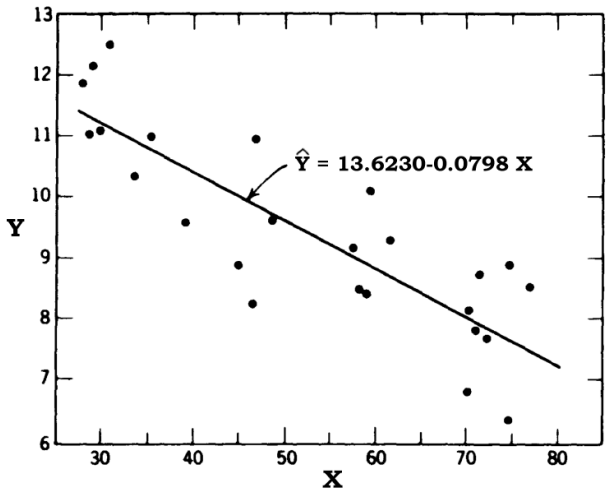


Figure 3.1: Scatter plot of observation y_i to x_i with fitted line and parameter estimates β_0 and β_1.

The terminology about fitting a line, parameters, a fitted model can be confusing. To fully understand what's the difference between the approximation of Y, what are estimated parameters, and what are unknown parameters, we can see Table 3.1.

Situation	θ	$\hat{\theta}$
Straight line model		
$Y = \beta_0 + \beta_1 X + \epsilon$	β_1	b_1
	β_0	b_0
Predicted response		
$\hat{Y} = b_0 + b_1 X$		

Table 3.1: Explanation of known and unknown parameters, estimations and estimators. The term θ refers to an unknown parameter, symbol "hat" like $\hat{\theta}$ is for estimation of unknown parameters.

3.1.2 Projection matrix

Theorem 3.2 *Suppose that* $\boldsymbol{X}$ *is* $n \times p$ *of rank* p*, so that* $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'}$*. Then the following holds:*

(i) $\boldsymbol{H}$ *and* $\boldsymbol{I_n} - \boldsymbol{H}$ *are symmetric and idempotent.*

(ii) $rank\ (\boldsymbol{I_n} - \boldsymbol{H}) = tr(\boldsymbol{I_n} - \boldsymbol{H}) = n - p.$

(iii) $\boldsymbol{HX} = \boldsymbol{X}.$

Matrix $\boldsymbol{H}$ is called *projection matrix*[1] or *hat matrix* because we obtain predictions of $\boldsymbol{Y}$ as $\hat{\boldsymbol{Y}} = \boldsymbol{HY}$. Let us also define matrix $\boldsymbol{M} = \boldsymbol{I_n} - \boldsymbol{H}$. Furthermore, residuals could be expressed as:

$$\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{HY} = \boldsymbol{MY}, \tag{3.6}$$

and graphically interpreted as in Fig. 3.2.

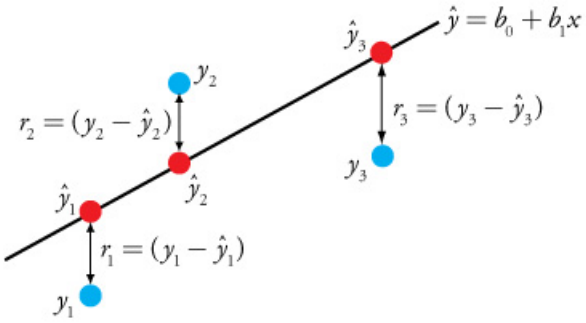


Figure 3.2: Plot of fitted model with interpretation of residuals.

3.1.3 Unbiased estimate of σ^2

Definition 3.3

$$RSS = (\boldsymbol{Y} - \hat{\boldsymbol{Y}})'(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{3.7}$$

is a residual sum of squares.

Residual sum of squares can be used to estimate the unknown σ^2.

Theorem 3.4 *If* $E(\boldsymbol{Y}) = \boldsymbol{X\beta}$, *where* $\boldsymbol{X}$ *is an* $n \times p$ *matrix of rank* r $(r \leq p)$, *and* $var(\boldsymbol{Y}) = \sigma^2 \boldsymbol{I_n}$, *then random variable*

$$s^2 = \frac{RSS}{n-r} = \frac{(\boldsymbol{Y} - \hat{\boldsymbol{Y}})'(\boldsymbol{Y} - \hat{\boldsymbol{Y}})}{n-r} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-r} \tag{3.8}$$

is an unbiased estimation of σ^2.

Proof. See [18] on page 45. □

Root square of unbiased estimation of variance (s) is called *residual standard error* and is used as summary characteristic for linear models.

3.2 Polynomial regression

The trend in data can have many shapes. Some scatter plots could show us trend that follows a polynomial function. If we set $x_{ij} = x_i^j$ and $k = p - 1$ $(\leq n - 1)$ in general multiple linear regression model, we obtain polynomial regression of kth-degree:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \epsilon_i \qquad (i = 1, 2, \ldots, n). \tag{3.9}$$

[1] In algebra, idempotent matrix is called projection matrix. It is because idempotent matrix $\boldsymbol{H}$ projects the observations of $\boldsymbol{Y}$ into a space generated by the columns of $\boldsymbol{X}$.

A very important term when speaking about polynomial regression is *orthogonal polynomials*. Model 3.9 would be transformed to formula

$$Y_i = \gamma_0\phi_0(x_i) + \gamma_1\phi_1(x_i) + \cdots + \gamma_k\phi_k(x_i) + \epsilon_i, \tag{3.10}$$

where $\phi_r(x_i)$ is an rth-degree polynomial in $x_i (r = 0, 1, \ldots, k)$, and orthogonality is described as follows:

$$\sum_{i=1}^{n} \phi_r(x_i)\phi_s(x_i) = 0 \qquad (allr, s, r \neq s). \tag{3.11}$$

How orthogonal polynomials work and what are their properties is subject to more complex regression analysis studies. Proper explanation can be found in [18].

3.3 Hypothesis testing

Since we can create many combinations of various variables, the need for testing a quality of the model and tests for comparing to other models arise. Several methods can be used for testing and evaluating the models, some of them stated below.

Let us assume a distribution of a random vector $\boldsymbol{X} = (X_1, \cdots, X_n)'$ that depends on parameter $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k)'$ and parameter $\boldsymbol{\theta}$ belongs to a set Ω, which we call *parametric space*. We don't know if $\boldsymbol{\theta}$ belongs to some nonempty subset ω of space Ω, therefore we call the statement $\boldsymbol{\theta} \in \omega$ a *null hypothesis*, noted $H_0 : \boldsymbol{\theta} \in \omega$. Opposite option is named *alternative hypothesis* and noted $H_1 : \boldsymbol{\theta} \notin \omega$. By the hypothesis test procedure, we decide about the H_0 based on the observations of random vector $\boldsymbol{X}$. Let $W \in B_n$ be a suitable set in $\mathbb{R}_n$ which we call *critical region*. In case of $\boldsymbol{X} \in W$ the null hypothesis $\boldsymbol{\theta} \in \omega$ is rejected. In case of $\boldsymbol{X} \notin W$ the null hypothesis isn't rejected. If the null hypothesis $\boldsymbol{\theta} \in \omega$ is rejected, in spite of being true, the conclusion is being of *type I error* (also known as "false positive", noted α). If the null hypothesis is not rejected, in spite of being false, the conclusion is of *type II error* (also known as "true negative", noted $1 - \beta$). Table 3.2 sums up decisions regarding the statistical errors.

	Null hypothesis H_0 is TRUE	Null hypothesis H_0 is FALSE
Reject null hypothesis	type I error (α)	correct decision
Fail to reject null hypothesis	correct decision	type II error ($1 - \beta$)

Table 3.2: Table of statistical errors that occur in hypothesis testing.

A common strategy for testing is to minimize the type II error with beforehand given parameter α. Usually, we choose $\alpha = 0.05$, but values 0.1 or 0.01 are also frequently used.

Till now there was no need for assumptions about the normality. For the rest of the chapter we will need that. Thus let us assume, that $\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ and therefore $\boldsymbol{Y} \sim N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$.

3.3.1 Variance stabilising transformation

If there are troubles ensuring the normality of residuals, we can use a transformation, so that the distribution of the transformed random variable would be very close to distribution fulfilling the assumptions of linear model.

Let us have a random variable X, whose distribution depends on some parameter θ. Assume $E(X) = \theta$. Usually, a variance of random variable X also depends on parameter θ, therefore $varX = \sigma^2(\theta)$. Let us find a function g, which is not constant and variable $Y = g(X)$ has variance, that would not depend on parameter θ. This issue usually doesn't have a solution, so let us find a suitable approximation. If g is function smooth enough, from a Taylor's polynomial we have

$$g(X) \doteq g(\theta) + (X - \theta)g'(\theta). \tag{3.12}$$

Therefore

$$Eg(X) \doteq g(\theta), \qquad var\, g(X) \doteq [g'(\theta)]^2\sigma^2(\theta). \tag{3.13}$$

If $g'(\theta)\sigma(\theta) = c$, where c is some constant, the expression $[g'(\theta)]^2\sigma^2(\theta)$ does not depend on θ. From this condition we get the result

$$g(\theta) = c \int \frac{d\theta}{\sigma(\theta)}. \tag{3.14}$$

The constant c is wisely chosen, so that the function g from (3.14) has suitable shape. Therefore $var\, g(X) \doteq c^2$. The function g is called *variance stabilising transformation*.

3.3.2 F-test

F-test is used to evaluate a significance of our regression. We want to test a hypothesis $H_0 : \boldsymbol{A\beta = c}$, where $\boldsymbol{A}$ is $q \times p$ of rank q and $\boldsymbol{\beta}$ is vector of length q. Natural testing statistics is $\boldsymbol{Ab = c}$ where $\boldsymbol{b}$ $(=\hat{\boldsymbol{\beta}})$ is our estimation of $\boldsymbol{\beta}$. Let us also define

$$\boldsymbol{b}_H = \boldsymbol{b} + (\boldsymbol{X'X})^{-1}\boldsymbol{A'}[\boldsymbol{A}(\boldsymbol{X'X})^{-1}\boldsymbol{A'}]^{-1}(\boldsymbol{c} - \boldsymbol{Ab}), \tag{3.15}$$

where $\boldsymbol{b}_H$ are the maximum likelihood estimates found using *Method of Lagrange Multipliers* (see [18], pages 60 and 99).

Then (3.7) becomes

$$RSS_H = (\boldsymbol{Y} - \hat{\boldsymbol{Y}}_H)'(\boldsymbol{Y} - \hat{\boldsymbol{Y}}_H) = ||\boldsymbol{Y} - \hat{\boldsymbol{Y}}_H||^2 = ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}_H||^2, \tag{3.16}$$

where RSS_H is the minimum value of $\boldsymbol{\epsilon'\epsilon}$ subject to $\boldsymbol{A\beta = c}$. Using eq. (3.15) with $\boldsymbol{c = 0}$ we have

$$\begin{aligned}
\hat{\boldsymbol{Y}}_H &= \boldsymbol{X}\hat{\boldsymbol{\beta}}_H \\
&= \left\{\boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'} - \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{A'}\left[\boldsymbol{A}(\boldsymbol{X'X})^{-1}\boldsymbol{A'}\right]^{-1}\boldsymbol{A}(\boldsymbol{X'X})^{-1}\boldsymbol{X'}\right\}\boldsymbol{Y} \\
&= (\boldsymbol{H} - \boldsymbol{H_1})\boldsymbol{Y} \\
&= \boldsymbol{H}_H\boldsymbol{Y},
\end{aligned}$$

where $\boldsymbol{H}_H$ is symetric and $\boldsymbol{H}_1$ is symmetric and idempotent (see [18] page 101).

An F-statistic for testing null hypothesis H_0 is described in following theorem.

Theorem 3.5

(i)

$$\begin{aligned}
RSS_H - RSS &= ||\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{\boldsymbol{H}}||^2 \\
&= (\boldsymbol{Ab} - \boldsymbol{c})'[\boldsymbol{A}(\boldsymbol{X'X})^{-1}\boldsymbol{A'}]^{-1}(\boldsymbol{Ab} - \boldsymbol{c}).
\end{aligned}$$

(ii)

$$E(RSS_H - RSS) = \sigma^2 q + (\boldsymbol{A\beta - c})'[\boldsymbol{A(X'X)^{-1}A'}]^{-1}(\boldsymbol{A\beta - c}) \\ = \sigma^2 q + (RSS_H - RSS).$$

(iii) When H is true,

$$F = \frac{(RSS_H - RSS)/q}{RSS/(n-p)} = \frac{(\boldsymbol{Ab-c})'[\boldsymbol{A(X'X)^{-1}A'}]^{-1}(\boldsymbol{Ab-c})}{qs^2},$$

then F has F-distribution $F(q, n-p)$ with q and n-p degrees of freedom.

(iv) When $\boldsymbol{c} = \boldsymbol{0}$, F can be expressed

$$F = \frac{n-p}{q}\frac{\boldsymbol{Y'(H-H_H)Y}}{\boldsymbol{Y'MY}}$$

where $\boldsymbol{H}_H$ is symmetric and idempotent, and $\boldsymbol{H}_H\boldsymbol{H} = \boldsymbol{H}\boldsymbol{H}_H = \boldsymbol{H}_H$.

Proof. See [18] page 100. □

F describes the ratio between s_H^2 and s^2 (where $s_H^2 = (RSS_H - RSS)/q$), so when we compare it to $100(1-\alpha)\%$ quantile of the $F(q, n-p)$ distribution we can determine if hypothesis H_0 is to be rejected or failed to be rejected.

When running F-test for one unknown parameter, F statistic is equal to second power of t statistic. This property is used for testing unknown parameters to be zero in linear models.

3.3.3 p-value

p-value is widely used for quick and easy interpretation of the outcomes of hypothesis testing. Most statistical softwares calculate it automatically with every hypothesis test. p-value represents the probability of obtaining same or greater observed value of our test statistic if the null hypothesis were true. For example, if we are doing an F-test from 3.3.2, first we calculate our F statistic that depends on our fitted model we are doing. Then the p-value would be understood as

Probability of (random variable with F(3,16) distribution > calculated F statistic) = p-value

For example, let us assume we obtained F-statistic = 2.23 with 3 and 16 degrees of freedom. Then

$$\mathrm{P(F} > 2.23) = 0.1241,$$

i.e. p-value equals 0.1241. To decide on null hypothesis H_0 we compare p-value to significance level α. If the p-value is less than α, we reject the null hypothesis because our F statistic is too "extreme". In our example, we can't reject the null hypothesis, because our F statistic is not "extreme" enough and p-value is greater than significance level.

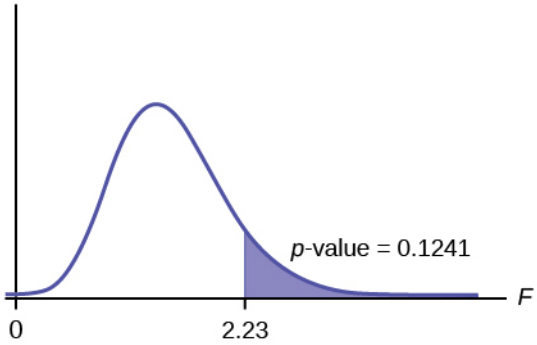


Figure 3.3: Graphical interpretation of example. Purple area is probability of F statistic appearing in this area when H_0 is true.

3.4 Model evaluation

Finding the best regression model to describe data can be a very tough and time consuming task. Relationships in natural world are very complex and many random factors come into consideration when trying to model the real case. To compare fitted models, evaluate the quality of models and therefore choose the one most suitable we use many tools out of which some are explained below.

3.4.1 AIC

Akaike information criterion is a type of criterion based on an idea of a discrepancy between the true distribution of $\boldsymbol{Y}$ which depends on $\boldsymbol{\theta}$ and the distribution specified by the model, which gives the estimation $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. Criterion is defined by following equation:

$$AIC = -2\log f(\boldsymbol{Y}; \hat{\boldsymbol{\theta}}(\boldsymbol{Y})) + 2r, \tag{3.17}$$

where $f(\boldsymbol{Y}; \boldsymbol{\theta})$ is the simultaneous density function of the random vector $\boldsymbol{Y}$, r is the dimension of the vector parameter $\boldsymbol{\theta}$. Further details can be found in [18].

For the purpose of our data analysis, the most important to understand is, that AIC doesn't work like hypothesis testing. It doesn't provide overall information about how good the fit is but tells us only how well is the model performing in comparison with other models. Smaller AIC value indicates better fit, so our objective is to minimize AIC value.

3.4.2 Coefficient of determination

Definition 3.6

$$R^2 = \frac{\sum(\hat{Y}_i - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2} = 1 - \frac{RSS}{\sum(Y_i - \overline{Y})^2} \tag{3.18}$$

is *coefficient of determination*.

R^2 represents the quantity of how much are predictors able to explain variability of response variable. The greater the R^2, the better the fit to observed data. If $Y_i = \hat{Y}_i$, we have perfect fit and therefore $R^2 = 1$, otherwise $R^2 < 1$. So basically our objective is to maximize R^2 when finding proper regression model or comparing models to each other.

3.4.3 Stepwise selection

Stepwise methods are methods of excluding explanatory variables in some manner with the aim of selecting the best subset from the initial basic model or its variations. Let us have a regression model (3.3)

$$\boldsymbol{Y} = \boldsymbol{X_p}\boldsymbol{\beta} + \boldsymbol{e},$$

where $\boldsymbol{X}$ is $n \times (k+1)$, and we want to identify the "significant" variables having nonzero regression coefficients. We divide k variables corresponding to the columnds of matrix $\boldsymbol{X_k}$ from (3.1) up into two sets: the first set consists of d variables that we regard as important, while the second set, which contains the remaining $k-d+1$ variables, consists of variables whose coefficient we suspect to be zero. We test if the regression parameters of the second set are equal to zero using F-statistic

$$F = \frac{RSS_d - RSS_{k+1}}{RSS_{k+1}} \frac{n-k-1}{k-d+1}, \tag{3.19}$$

where RSS_d, RSS_{k+1} are residual sums of squares of models consisting of indexed explanatory variables.

Backward elimination – is started with full model using all k variables and compute (3.19) with $d = k$ for each of the k variables. We eliminate the variable having the smallest F-statistic from the model. We continue this procedure until all variables are eliminated or until p-value of the test is smaller than chosen significance level.

4 Sample data analysis

4.1 Introduction

With development of technology and implementing it to people's everyday lives, humankind's dependence on energy grows. Thus, search for sustainable and environmentally friendly production of huge amount of energy takes important role in science research. One of the efficient sources of energy is nuclear process, especially uranium fission. uranium can be formed in sandstone-hosted uranium deposits, what accumulates approximately 18% of world's known reserves of Uranium. Study of sandstone-hosted uranium, its ore mineralization, mineral phases, gel nature of components can help us understand uranium interactions with other elements and enhance our capabilities of detecting uranium in nature [14]. For the analysis we are using R programming language on data obtained by LIBS in laboratory of BUT.

4.2 Problem introduction

4.2.1 Sample

The analysed sample was taken from Břevniště deposit in the area of northern part of the Bohemian Cretaceous Basin in Czech Republic. It was cut, dried, and cemented in Araldite epoxy. Then flattened surface was achieved by grinding the excess of epoxy and cropping edges. All of this is requirement of X-ray Fluorescence (XRF) analysis. In Fig. 4.1 we can see the XRF analysis of the sample.

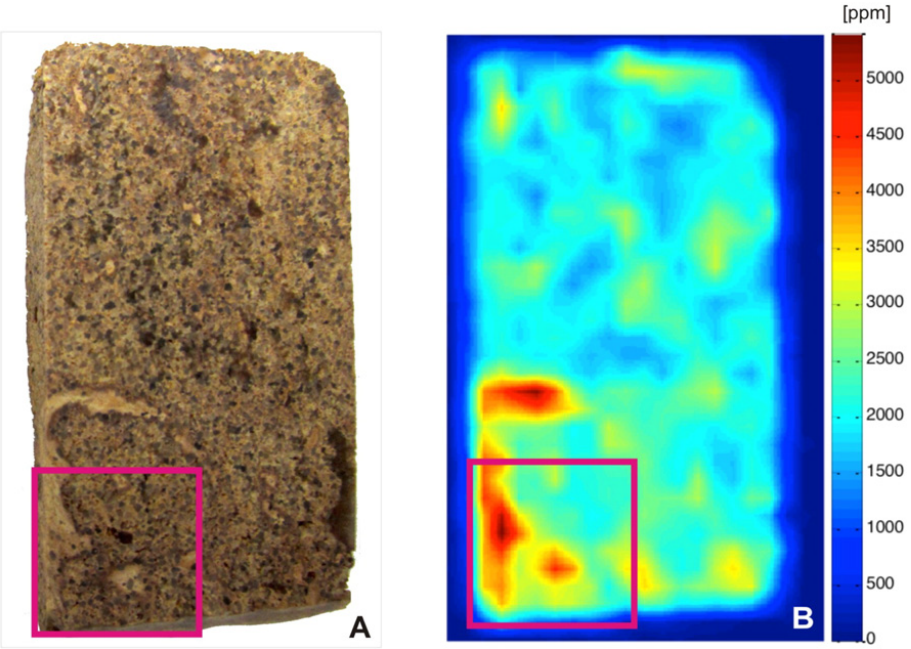
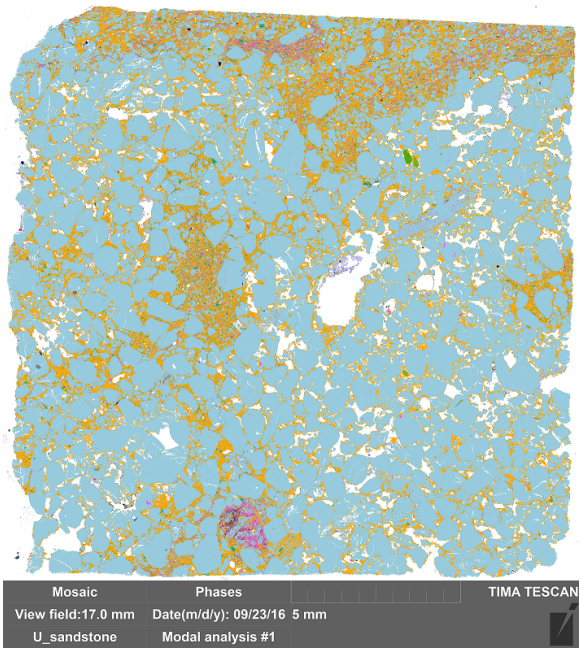


Figure 4.1: Maps of spatial distribution of selected elements in the sample of uranium ore provided using XRF analysis. A) photography of scanned sample (sized 70×44mm); B) distribution of uranium content within the sample surface. Red square highlights the region for further LIBS analysis (15 × 15 mm) [14].

4.2.2 Geological interpretation

Silicon (Si) and zirconium (Zr) are minerally bonded to phase called as "hydrozirconium" ($ZrSiO_4 \cdot nH_2O$). The ore sample – sandstone-hosted uranium – contains a lot of mineral phases, as seen in Fig 4.2. The point of interest is to describe the relation between uranium and hydrozirconium. We are trying to answer how does uranium interact individually with either zirconium and silicon, if there is a statistical inference between those two elements and if it is important for explanation of presence of uranium.



(a) Pixel map

(b) Legend

Figure 4.2: Chemical anlalysis of sample using QEMSCAN, a method developed for revealing ore mineralization based on combination of energy-dispersive X-ray spectroscopy and backscattered electron measurements in scanning electron microscope.

A pixel map, where every pixel is coloured by the most probable mineral is available for reference with the LIBS measurement.

4.3 LIBS data representation

Same sample was analyzed by LIBS apparatus at BUT. We did 150 ×150 measurements, out of which were obtained 22 500 emission spectra, 3 907 800 intensities in 4.5 GBs of data. Each spectrum is measured in interval 198.7959 nm - 1016.77 nm wavelength, divided to non-constant spaces. From every spectrum we separated corresponding intensities for every important chemical element located in the sample. All elements and notation representing each element for programming are listed in Tab. 4.1.

Element	Symbol	Notation
Uranium	U	u
Silicon	Si	s
Zirconium	Zr	z
Niobium	Nb	nb
Aluminium	Al	a1, a2
Ferrum	Fe	f1, f2
Hafnium	Hf	h1, h2
Titanium	Ti	t1, t2

Table 4.1: Notation of elements used in programming.

Data for every element were evaluated as sum of intensities in an interval of length 0.2 centered at certain wavelength corresponding to emission lines of elements based on NIST database [2]. Therefore analysed intensities of elements are sums of 6 to 13 values (intensities) in the original spectrum within given interval. As mentioned in 2.5, theoretically it should be just one point for every element. In reality the line is broadened by several mechanisms. Mathematical models are tested on transformations of data, i.e. standardisation[1] and local standardisation[2].

Best model fits and meeting assumptions of linear model are reached when using detected "raw" data with no transformation. Because absolute intensity of a selected spectral line is not as important as proportions between intensities of lines, we divided all intensities by 10^6 for clearer notation. Relevance of the data remained the same. Some elements are present at two wavelengths in the spectrum. That is because in reality, all elements are represented by multiple spectral lines. Individual lines of respective element behave differently with large changes of chemical composition. Selecting multiple lines helps to develop more complex model that can improve stability with regards to spectral interference. For our sample we chose wavelengths based on empirical experience. Thus to keep reliability of the initial model we had to observe some elements present at two wavelengths.

Raster maps of four elements measured with LIBS with wavelengths are shown in Fig. 4.3. Maps of eight others are attached in attachment section.

[1] Before extracting emission lines of the elements, every value in the spectrum is subtracted by average of all spectral intensities and divided by standard deviation of whole spectrum.

[2] When taken sum of 0.2 length interval, sum is divided by standard deviation of those 6 to 13 points summed.

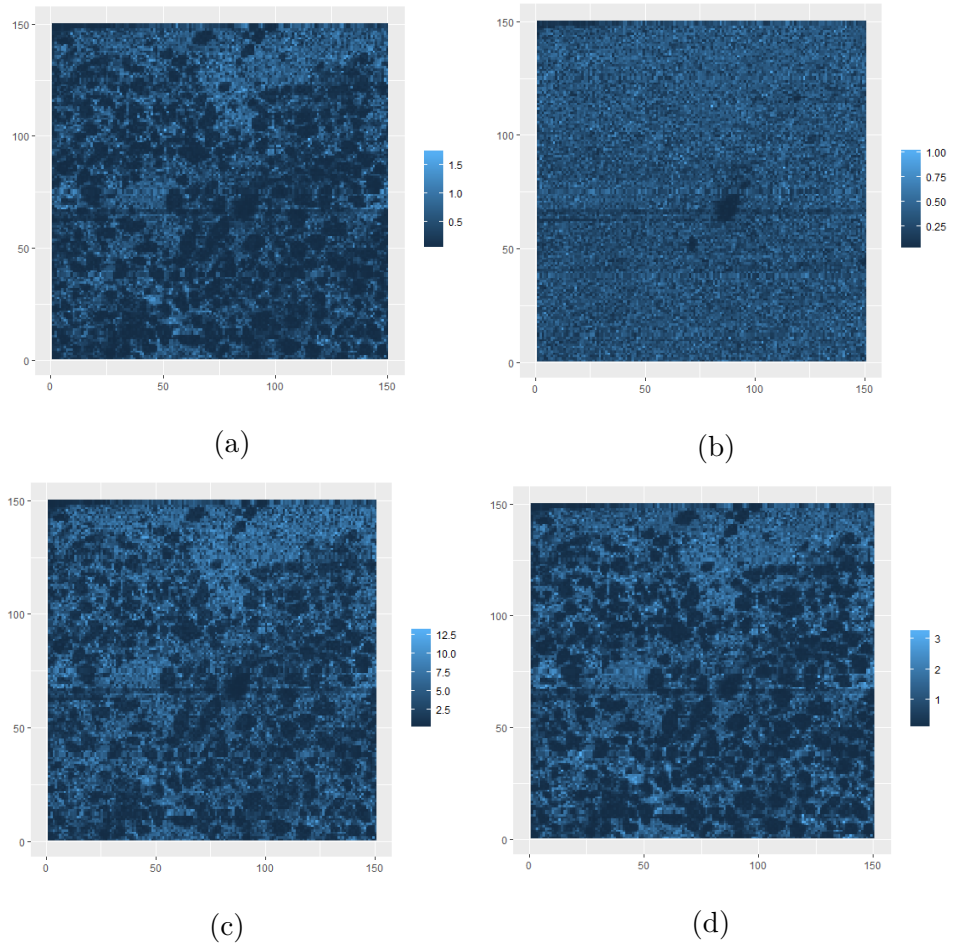


Figure 4.3: Raster map of elements measured at wavelengths: (a) uranium - 409.02 nm, (b) Silicon - 251.431 nm, (c) Zirconium - 349.621 nm, (d) Niob - 405.89 nm.

As we can see, elements that belongs to common mineral phase highly correlate in the raster maps. The comparison of LIBS and QEMSCAN outputs also shows a good correspondence, however LIBS map denotes relative abundance and QEMSCAN reveals individual mineralization phases.

4.4 Modeling

Initial problem is to explain intensity of uranium by other predictive elements. Therefore basic model all our modeling came from is *polynomial regression model of 3rd degree with paired interferences*. Geological studies postulated a high influence of zirconium and silicon on the presence of uranium. Thus further in the study we are especially focusing on this particular relation. For all tests we use 5% significance level ($\alpha = 0.05$). In R code the basic model is represented as formula like this:

$$
\begin{aligned}
u \quad \sim \quad & s + I(s^2) + I(s^3) + \\
& z + I(z^2) + I(z^3) + \\
& a1 + I(a1^2) + I(a1^3) + \\
& a2 + I(a2^2) + I(a2^3) + \\
& f1 + I(f1^2) + I(f1^3) + \\
& f2 + I(f2^2) + I(f2^3) + \\
& h1 + I(h1^2) + I(h1^3) + \\
& h2 + I(h2^2) + I(h2^3) + \\
& nb + I(nb^2) + I(nb^3) + \\
& t1 + I(t1^2) + I(t1^3) + \\
& t2 + I(t2^2) + I(t2^3) + \\
& s:z + s:a1 + s:a2 + s:f1 + s:f2 + s:h1 + s:h2 + s:nb + \\
& s:t1 + s:t2 + \\
& z:a1 + z:a2 + z:f1 + z:f2 + z:h1 + z:h2 + z:nb + \\
& z:t1 + z:t2 + \\
& a1:a2 + a1:f1 + a1:f2 + a1:h1 + a1:h2 + a1:nb + \\
& a1:t1 + a1:t2 + \\
& a2:f1 + a2:f2 + a2:h1 + a2:h2 + a2:nb + a2:t1 + a2:t2 + \\
& f1:f2 + f1:h1 + f1:h2 + f1:nb + f1:t1 + f1:t2 + \\
& f2:h1 + f2:h2 + f2:nb + f2:t1 + f2:t2 + \\
& h1:h2 + h1:nb + h1:t1 + h1:t2 + \\
& h2:nb + h2:t1 + h2:t2 + \\
& nb:t1 + nb:t2 + \\
& t1:t2,
\end{aligned}
\tag{4.1}
$$

where $I(x^p)$ converts variable to power p (necessary condition for creating polynomial models in R) and $x_i : x_j$ means interaction between variables x_i and x_j. We used linear model function in R called `lm()`. Basic model contains 88 terms, out of them 54 are not significant according to their p-values. F test of the model is significant and coefficient of determination (R^2) is 98.55%. Residuals are heteroscedastic and don't appear to be normally distributed. We consider basic model to be inappropriate to explain intensity of uranium.

Afterwards, we tried transformations of response and explanatory variables ($ln(u), u^2$, $\sqrt{u}, \frac{1}{u}$ and other combinations of formulas. Models on standardised and locally standardised data were also tested.

4.5 Suitable model

The most suitable model was obtained utilizing **raw data** (no standardisation) and **transformed response** $\sqrt{u}$. Residuals are homoscedastic and appear normal even in basic model. We created submodel by backward stepwise regression, i.e. by excluding the most insignificant term (highest p-value) and comparing the obtained submodel with using F test - calculated by `anova()` R function. We proceeded this till result of anova test rejected that the full model can be reduced to the last model. Resulted model has formula:

$$\begin{aligned} sqrt(u) \quad \sim \quad & s + I(s^2) + z + I(z^2) + I(z^3) + \\ & nb + I(nb^2) + I(nb^3) + \\ & a1 + I(a1^2) + I(a1^3) + a2 + I(a2^3) + \\ & f1 + I(f1^2) + I(f1^3) + f2 + I(f2^2) + I(f2^3) + \\ & h1 + I(h1^3) + h2 + I(h2^2) + I(h2^3) + \\ & t1 + I(t1^2) + I(t1^3) + t2 + I(t2^2) + I(t2^3) + \\ & s:z + s:f1 + s:f2 + s:h1 + s:nb + \\ & z:a2 + z:f1 + z:f2 + z:h1 + z:t2 + \\ & a1:a2 + a1:h2 + a2:f1 + a2:f2 + a2:h1 + a2:h2 + a2:nb + \\ & f1:h2 + f1:nb + f1:t1 + h1:h2 + h1:t1 \end{aligned} \tag{4.2}$$

The parameter estimates together with their standard error, value of the T statistic and its p-value are given in table 4.2:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.256e-01	1.153e-03	1.090e+02	0.000e+00
s	-1.140e-02	6.670e-03	-1.709e+00	8.743e-02
I(s^2)	6.251e-02	1.003e-02	6.234e+00	4.642e-10
z	1.113e-01	1.902e-03	5.850e+01	0.000e+00
I(z^2)	-7.574e-03	4.861e-04	-1.558e+01	1.943e-54
I(z^3)	3.570e-04	2.371e-05	1.505e+01	5.728e-51
a1	1.039e-01	4.067e-02	2.553e+00	1.067e-02
I($a1^2$)	-1.417e+00	3.837e-01	-3.693e+00	2.222e-04
I($a1^3$)	2.738e+00	3.458e-01	7.920e+00	2.492e-15
a2	6.224e-02	1.032e-02	6.031e+00	1.653e-09
I($a2^3$)	-1.030e-02	2.782e-03	-3.703e+00	2.138e-04
f1	-1.628e-01	1.062e-02	-1.533e+01	8.407e-53
I($f1^2$)	5.298e-02	8.449e-03	6.270e+00	3.676e-10
I($f1^3$)	-1.263e-02	1.182e-03	-1.068e+01	1.400e-26
f2	3.362e-02	3.937e-03	8.540e+00	1.428e-17
I($f2^2$)	-1.511e-02	1.041e-03	-1.452e+01	1.557e-47
I($f2^3$)	1.046e-03	7.853e-05	1.332e+01	2.506e-40

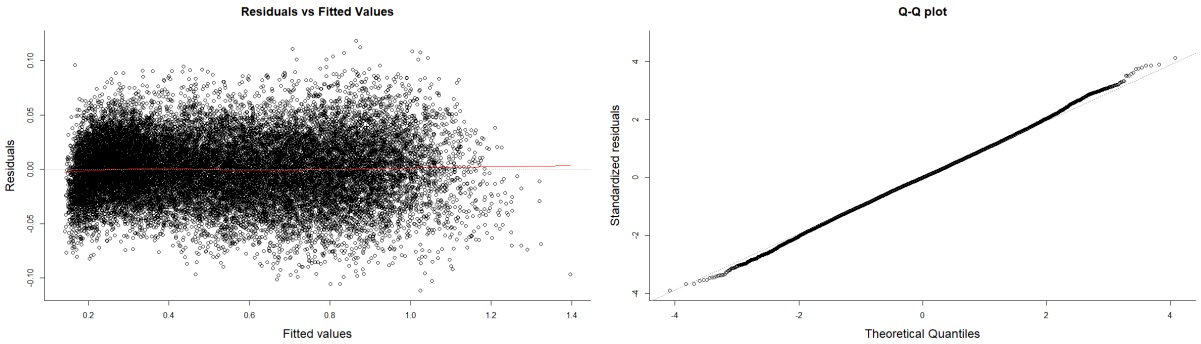
h1	5.454e-03	2.145e-02	2.543e-01	7.993e-01
I($h1^3$)	-8.931e-02	1.782e-02	-5.013e+00	5.389e-07
h2	6.477e-01	5.660e-02	1.144e+01	3.091e-30
I($h2^2$)	-3.919e+00	6.790e-01	-5.772e+00	7.929e-09
I($h2^3$)	6.228e+00	9.996e-01	6.230e+00	4.744e-10
nb	2.404e-01	7.502e-03	3.204e+01	2.376e-220
I(nb^2)	-5.242e-02	5.912e-03	-8.867e+00	8.076e-19
I(nb^3)	7.930e-03	1.101e-03	7.204e+00	6.046e-13
t1	4.402e-02	1.344e-02	3.275e+00	1.057e-03
I($t1^2$)	-3.068e-01	2.586e-02	-1.186e+01	2.318e-32
I($t1^3$)	7.999e-02	9.888e-03	8.090e+00	6.268e-16
t2	2.607e-01	5.270e-02	4.946e+00	7.620e-07
I($t2^2$)	-1.045e+00	5.867e-01	-1.781e+00	7.488e-02
I($t2^3$)	3.849e+00	1.352e+00	2.847e+00	4.417e-03
s:z	-5.162e-02	3.270e-03	-1.579e+01	7.746e-56
s:f1	2.067e-01	2.067e-02	1.000e+01	1.631e-23
s:f2	-6.875e-02	7.268e-03	-9.460e+00	3.381e-21
s:h1	3.208e-01	4.476e-02	7.167e+00	7.911e-13
s:nb	-4.328e-02	1.227e-02	-3.527e+00	4.213e-04
z:a2	-1.114e-02	2.649e-03	-4.206e+00	2.614e-05
z:f1	-1.030e-02	3.282e-03	-3.139e+00	1.699e-03
z:f2	1.156e-02	1.027e-03	1.125e+01	2.861e-29
z:h1	-1.634e-02	5.102e-03	-3.202e+00	1.367e-03
z:t2	-6.959e-02	9.432e-03	-7.379e+00	1.656e-13
a1:a2	-2.746e-01	8.869e-02	-3.096e+00	1.961e-03
a1:h2	1.805e+00	2.904e-01	6.217e+00	5.147e-10
a2:f1	8.387e-02	1.407e-02	5.959e+00	2.576e-09
a2:f2	-3.016e-02	4.552e-03	-6.626e+00	3.530e-11
a2:h1	1.808e-01	3.046e-02	5.936e+00	2.969e-09
a2:h2	-2.195e-01	8.987e-02	-2.443e+00	1.458e-02
a2:nb	5.468e-02	6.681e-03	8.184e+00	2.902e-16
f1:h2	1.832e-01	6.008e-02	3.050e+00	2.293e-03
f1:nb	-3.483e-02	5.851e-03	-5.953e+00	2.664e-09
f1:t1	5.721e-02	8.665e-03	6.603e+00	4.125e-11
h1:h2	-4.078e-01	2.123e-01	-1.921e+00	5.480e-02
h1:t1	1.799e-01	2.799e-02	6.426e+00	1.336e-10

Table 4.2: Summary of submodel with regression coefficients (estimates) and p-values using R function `summary()`.

Residual standard error	0.02883 on 22447 degrees of freedom
Multiple R-squared	0.9876
F-statistic	3.441e+04 on 52 and 22447 DF
p-value	$<$ 2.2e-16

Table 4.3: Test of whole submodel – F test and coefficient of determination.

Plot of residuals and Quantile-Quantile plot are shown below.



(a) Residual plot (b) Q-Q plot

Figure 4.4: Graphical assessment of the regression assumptions. In (*a*) a red line is a spline constructed over residuals and dotted line in (b) plot represents best data fit to normal distribution.

Fig. 4.4 shows us that visually residuals are homoscedastic, e.g. variance is constant, and are also very similar to quantiles of normal distribution, therefore appear as normally distributed. Among all the other models tested these assumptions were met in this submodel the most.

F test for initial basic model and proceeded submodel shows us, that we can't reject that the basic model can be reduced to the submodel.

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	22406	18.55				
2	22442	18.58	-36	-0.03	1.02	0.4352

Table 4.4: Analysis of variance table for basic model (1) vs. submodel (2).

P-value is greater than 0.05, therefore hypothesis can't be rejected.

The prediction capabilities of our fitted model can be assessed in Fig. 4.5.

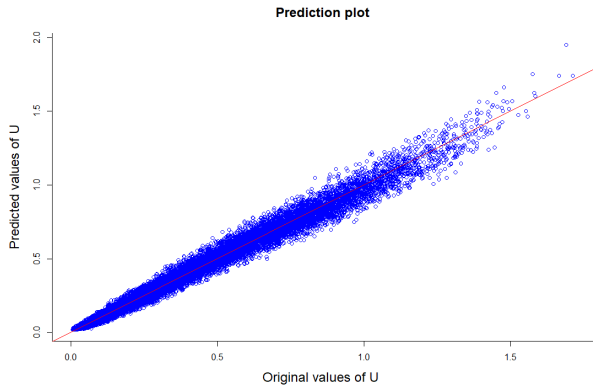


Figure 4.5: Prediction of U based on suitable submodel

4.6 Influence of silicon and zirconium

In the summary from Tab. 4.2 we can observe presence of silicon to 1st and 2nd power, zirconium to 1st, 2nd and 3rd power, and inference between zirconium and silicon. All of these terms are significant except 1st power of silicon, which is acceptable in complex models like this, because silicon plays role in inferences with other elements that are statistically significant, therefore we can not exclude it from the model. When we consider all the other terms to be constant and let only Zr and Si change, we obtain estimated model:

$$\tilde{U}_{SZ} = -0.0114S + 0.0625S^2 + 0.1113Z - 0.0076Z^2 + 0.0004Z^3 - 0.0516ZS \qquad (4.3)$$

Using the `mesh()` function from MATLAB, we can represent the uranium predictions with respect to changes of zirconium and silicon:

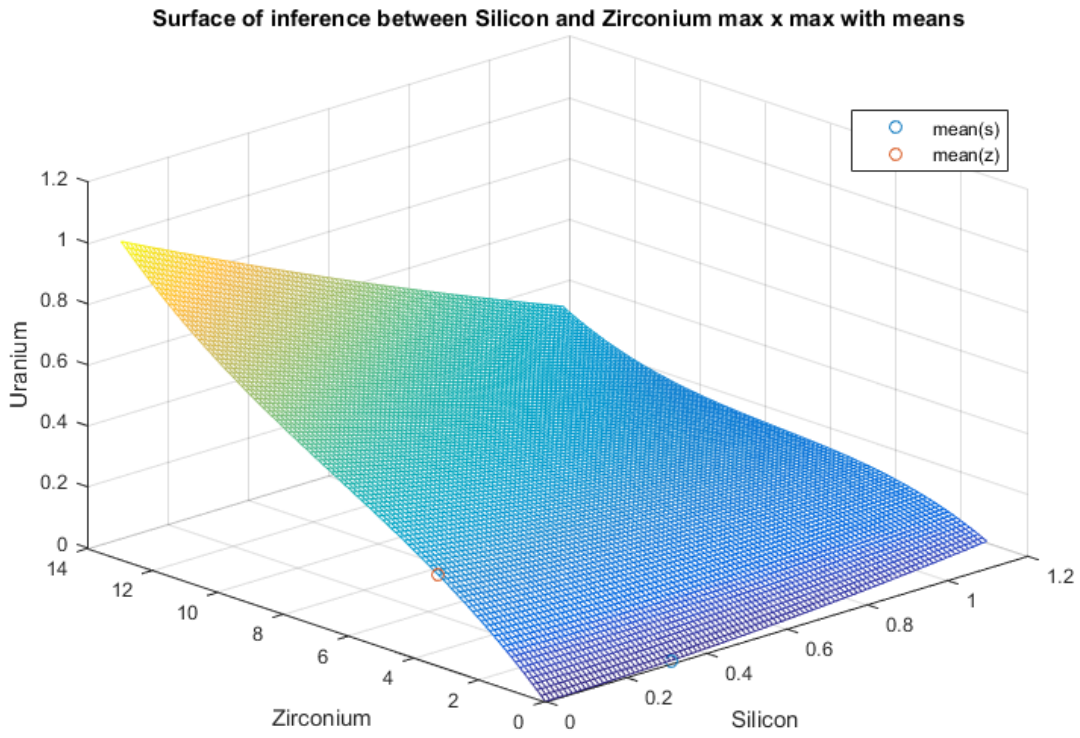


Figure 4.6: Surface of inference between Si and Zr with the result in U under the assumption that other inferences are zero. The limits of the axes are set to minimum and maximum values of Si and Zr with blue and red dot as means of covariates.

We can observe how zirconium appears to have much bigger influence on uranium than silicon. Also if we observe trend when Si and Zr are simultaneously rising, it affects also rise of U. Therefore graphically we can confirm inference between Zr and Si in our sample. Important thing to mention here is, that proportion of axes in Fig. 4.6 is very big. Length of axes is based on minimum and maximum value observed in the data. Consequently, for our sample U seems to be influenced by Zr more than by Si. But as is mentioned in 4.3, the important information is yielded by proportions between intensities not by the absolute value of the intensity. Thus if we prolong the length of axes equally, we get much different result and inference trend shown in Fig. 4.7

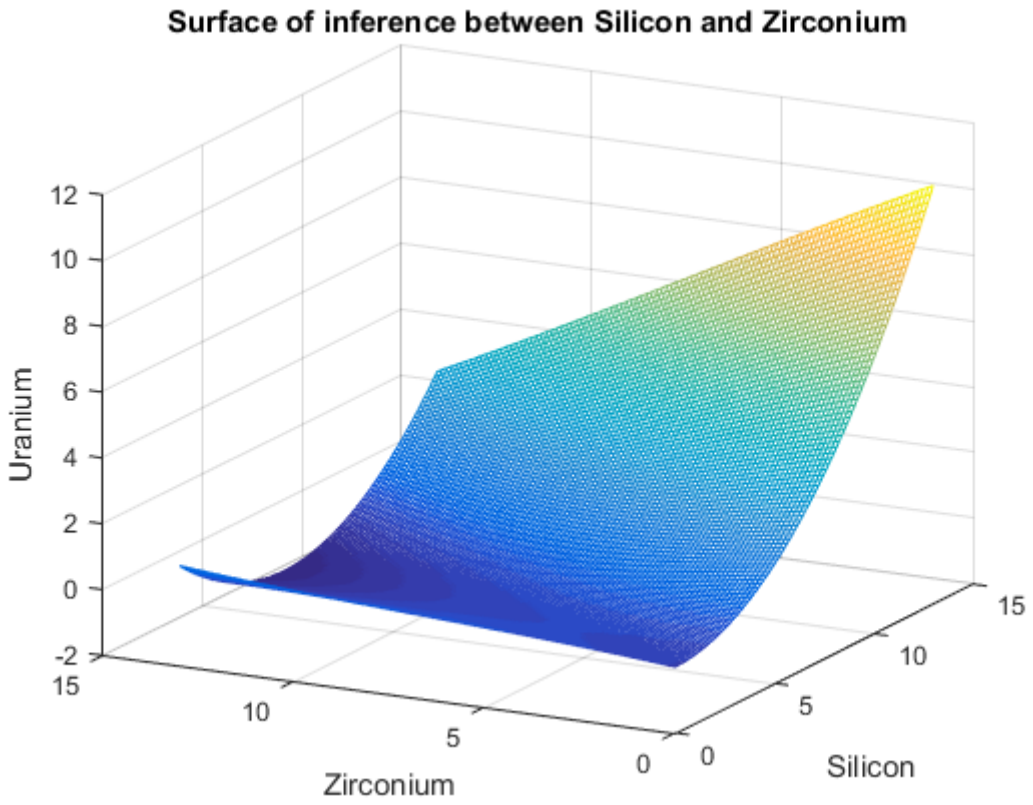


Figure 4.7: Surface of inference between Si and Zr with equal range on x and y axis.

The applicability of model in high Si intensities or under different assumptions about interferences should be a subject to further studies.

4.7 Comparison with generalized linear model

For the purpose of checking our deduction about suitable submodel we tried one more approach – generalized linear model [6]. We established R model glm(u ., family = Gamma(link = „sqrt“)) that is equal to our basic linear model. Then using the stepwise regression we excluded insignificant terms and found the sufficient glm submodel. Sufficient glm submodel seems to be worse than lm submodel, regarding heteroscedastic residuals and not the sufficient normality fit of residuals[3]. Tests and prediction capabilities are stated in attachments. When we compare which terms appear in glm submodel and lm submodel, we can detect 17 out of 22 inferences to appear in glm model as well as in lm model, what supports suitability of lm model. In Fig. 4.8 we can see prediction difference between lm and glm model.

For lower values glm tends to predict higher values than lm model, at intensities of value 1 trend changes and lm model predicts higher values than glm model. Another interesting point to realize is, that glm model is also unable to predict lower values, since lm starts at level close to zero, glm starts at values of 0.2.

[3] Anscombe transformation transforms residuals to be homoscedastic and normal.

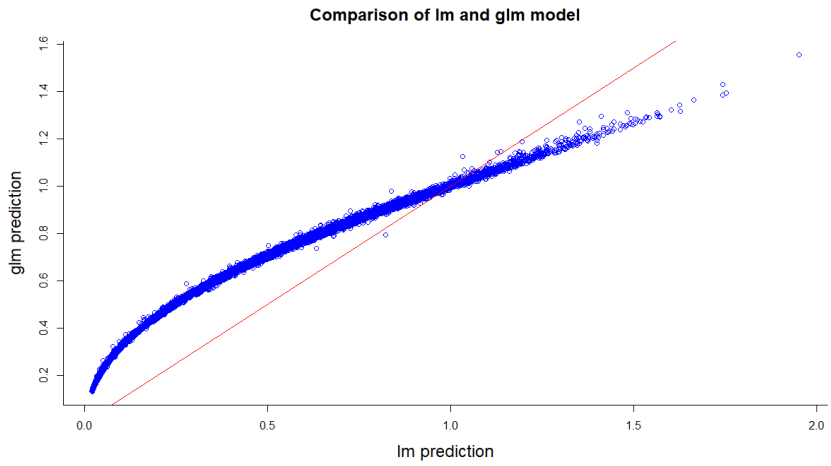


Figure 4.8: Comparison of linear model and generalized linear model. Red line is where are lm and glm equal.

4.8 Discussion

From the analysis we can say, that using ”raw”, not transformed data during pre-processing phase is the best ground for modeling. Transformations of the collected data deteriorate results, theoretical assumptions, and abilities of models to explain real relations.

On the other hand, during modeling transformation of response variable U is highly demanding. Best transformation of response we found is square root of a random variable. Compared to other transformations of variables, model with square root of U and non-transformed explanatory variables resulted in homoscedastic, almost normally distributed residuals with constant variance, which is not met with other transformations.

Suitable linear submodel has 11 explanatory variables (7 chemical elements) up to 3rd power, 52 regression terms, 22 inferences, and 45 terms are statistically significant in the model. Coefficient of determination for this model is 98.76%.

Elements silicon and zirconium are both significant in the model. Their inference is also statistically significant, thus we can not terminate either element from the model on explaining presence of uranium. Therefore phase hydrozirconium seems to be significant and both elements of this phase play important role in the sample.

5 Conclusion

The goal of this thesis was to search for the most suitable statistical model for explaining presence of uranium by presence of other chemical elements in sandstone-hosted uranium deposit in Břevniště. Also to explore and describe a relation between uranium, and zirconium and silicon as constituents of hydrozirconium.

In chapter 2 we introduced the analytical technique called LIBS. We explained what is the basic principle of this spectroscopic method, what happens when we use laser light as a source of energy for AES, how the radiation is detected, how the data is obtained, and how do apparatus look in the laboratory at BUT. Understanding how physical processes work is essential for further statistical analysis, especially because of complex statistical modeling.

In chapter 3 we set the basic knowledge about regression analysis used for analysing our sample. We showed basic regression formula which was also used for our sample. We explained basic evaluation tools crucial for deciding, whether model is appropriate, and comparing tools to finding the most suitable model.

In chapter 4 we used the basis established in chapters 2 and 3, and built the model on these fundamentals. In the beginning, we explained the basic problem and its importance for research. We described sample itself, its origin, how it was processed and analysed with different tools except LIBS. Then we explained how the most suitable linear regression model was found, and we stated its properties and plots. We applied discovered model to our initial problem and explored its behaviour. After all that, we tried to support our conclusions with establishing glm model, which we used to compare the results.

References

[1] ANDĚL, J. *Základy matematické statistiky.*, Praha: Matfyzpress, 2011. ISBN 978-80-7378-1620.

[2] Atomic Spectra Database. *National Institute of Standards and Technology: Physical measurement laboratory* [online]. Last revision 3rd of November 2017 [cit. 2018-05-03], <https://www.nist.gov/pml/atomic-spectra-database>.

[3] *Automated 2D elemental mapping by LIBS* [image], <http://www.andor.com/learning-academy/automated-2d-elemental-mapping-by-laser-induced-breakdown-spectroscopy-application-note> [accessed 2018-4-7].

[4] BARNES, R. M. *Emission Spectroscopy.* Stroudsburg, PA: Dowden, Hutchison & Ross, 1973.

[5] CIUCCI, A., et al. New Procedure for Quantitative Elemental Analysis by Laser-Induced Plasma Spectroscopy. *Applied Spectroscopy* [online]. 1999, vol. 53, no. 8, p. 960–964 [cit. 2018-04-24]. DOI: 10.1366/0003702991947612.

[6] DOBSON, A. J. *An introduction to generalized linear models.* 2nd ed. Boca Raton: Chapman & Hall, c2002. Chapman & Hall texts in statistical science series. ISBN 1-58488-165-8.

[7] DRAPER, N. R., SMITH, H. *Applied regression analysis.* 3rd ed. New York: Wiley, c1998. ISBN 0-471-17082-8.

[8] *Emission of Continuum* [image]. <https://appliedspectra.com/technology/libs.html> [cit. 2018-3-29]

[9] GALTON, F. Regression Towards Mediocrity in Hereditary Stature. *Journal of the Anthropological Institute of Great Britain and Ireland.* 1886, vol. 15, 246-263 [cit. 2018-05-17]. DOI: 10.2307/2841583, <http://www.jstor.org/stable/2841583>.

[10] HADDAD, J. E., CANIONI, L., BOUSQUET, B. Good practices in LIBS analysis: Review and advices. *Spectrochimica Acta Part B: Atomic Spectroscopy* [online]. 2014, vol. 101, p. 171-182 [cit. 2018-04-28]. DOI: 10.1016/j.sab.2014.08.039, <http://linkinghub.elsevier.com/retrieve/pii/S0584854714002158>. ISSN 05848547.

[11] HAHN, D. W., OMENETTO N. Laser-Induced Breakdown Spectroscopy (LIBS), Part I: Review of Basic Diagnostics and Plasma—Particle Interactions. *Applied Spectroscopy* [online]. 2010, vol. 64, no. 12, 335A-336A [cit. 2017-09-07]. DOI: 10.1366/000370210793561691. ISSN 0003-7028.

[12] HUTCHINSON, I. H. *Principles of plasma diagnostics.* 2nd ed. Cambridge: Cambridge University Press, 2002. ISBN 0-521-80389-6.

[13] KIRCHOFF, G., BUNSEN, R. *Chemische Analyse durch Spectralbeobachtungen.* Wien: Verl. Fabrik u. handlung, 1860.

[14] KLUS J., et al. Multivariate approach to the chemical mapping of uranium in sandstone-hosted uranium ores analyzed using double pulse Laser-Induced Breakdown Spectroscopy. *Spectrochimica Acta Part B: Atmoic Spectroscopy* [online]. 2016, vol. 123, p. 143-149 [cit. 2018-5-10]. DOI: 10.1016/j.sab.2016.08.014, <http://www.sciencedirect.com/science/article/pii/S058485471630146X>. ISSN 0584-8547.

[15] LAMOŠ, F., POTOCKÝ, R. *Pravdepodobnosť a matematická štatistika: štatistické analýzy.* 2nd ed. Bratislava: Univerzita Komenského, 1998. ISBN 80-223-1262-2.

[16] MIZIOLEK, A. W., PALLESCHI, V., SCHECHTER, I. *Laser-induced breakdown spectroscopy (LIBS): fundamentals and applications.* New York: Cambridge University Press, 2006. ISBN 978-0-521-85274-6.

[17] NOVOTNÝ, J. Dálkově řízená laserová spektroskopie (LIBS). Brno: Vysoké učení technické v Brně, Fakulta strojního inženýrství, 2012. 97 p. Supervisor doc. Ing. Jozef Kaiser, Ph.D.

[18] SEBER, G. A. F., LEE, A. J. *Linear regression analysis.* 2nd ed. Hoboken, N.J.: Wiley-Interscience, c2003. ISBN 0-471-41540-5.

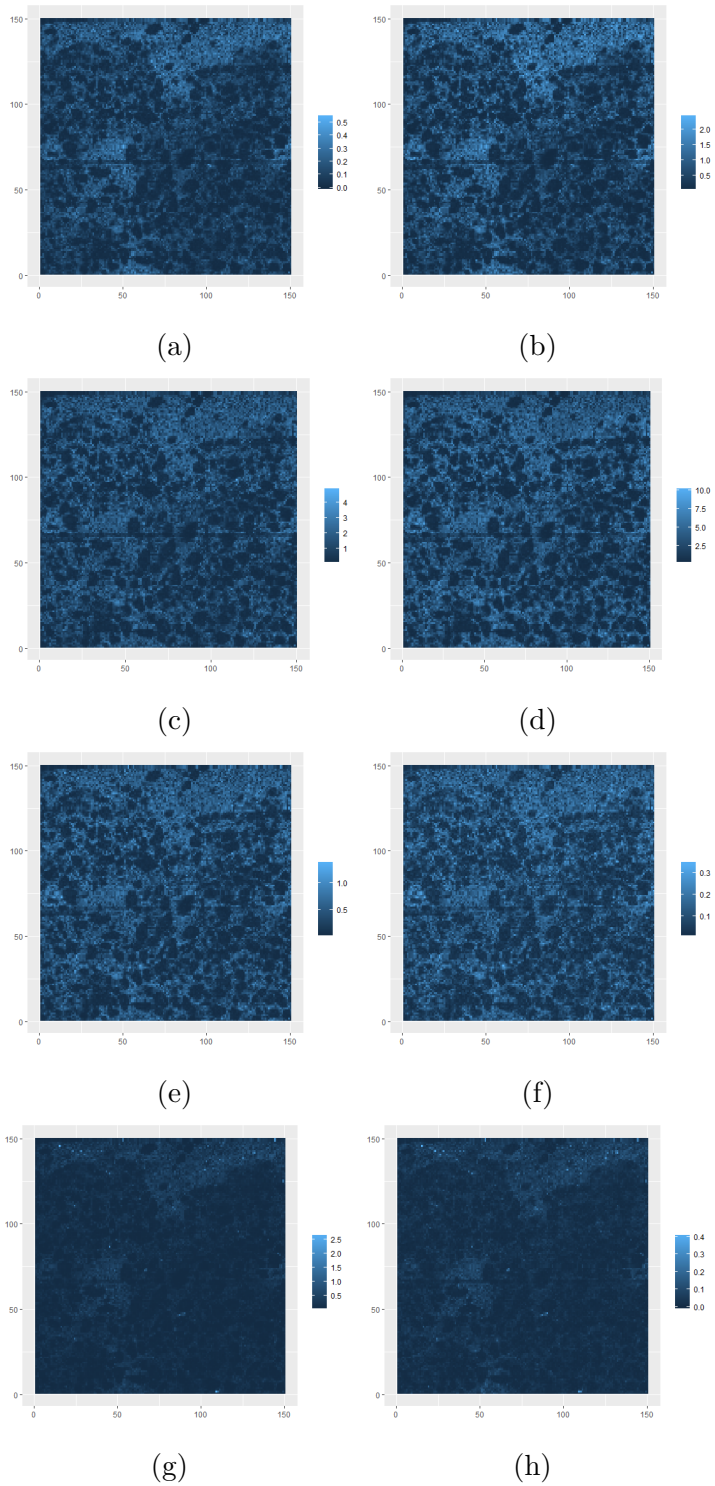
[19] TOROK, T., MIKA, J. M., GEGUS, E., *Emission Spectrochemical Analysis.* Bristol: Adam Hilger, 1978.

[20] VOGEL, A., VENUGOPLAN, V. Mechanisms of Pulsed Laser Ablation of Biological Tissues. *Chemical reviews* [online]. 2003, vol. 103, no. 2, pp. 577-644 [cit. 2018-04-10]. DOI: 10.1021/cr010379n, <http://pubs.acs.org/doi/abs/10.1021/cr010379n>. ISSN 0009-2665.

Used symbols

$\boldsymbol{\beta}$	vector of unknown regression parameters
λ_e	wavelength at which element emits radiation
σ^2	variance of Normal distribution
θ	unknown parameter
X, Y	random variable
$\boldsymbol{X, Y}$	vector/matrix
$\boldsymbol{I_n}$	identity matrix of size n, $n \times n$
$\boldsymbol{X'}$	transpose of vector/matrix X
$\boldsymbol{X^{-1}}$	inverse matrix
$\hat{\boldsymbol{Y}}$	predictions
$tr(X)$	trace of matrix X
$E(X)$	expected value of random variable
$varX$	variance of random variable
$N(\mu, \sigma^2)$	Normal distribution
$F(p, q)$	F distribution

Attachments



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 5.1: Raster map of rest of the elements used in modeling with their wavelengths: (a) Aluminium 1 - 308.24 nm, (b) Aluminium 2 - 309.31 nm, (c) Ferrum 1 - 302.06 nm, (d) Ferrum 2 - 404.58 nm, (e) Hafnium 1 - 368.24 nm, (f) Hafnium 2 - 417.46 nm, (g) Titanium 1 - 323.50 nm, (h) Titanium 2 - 498.17 nm

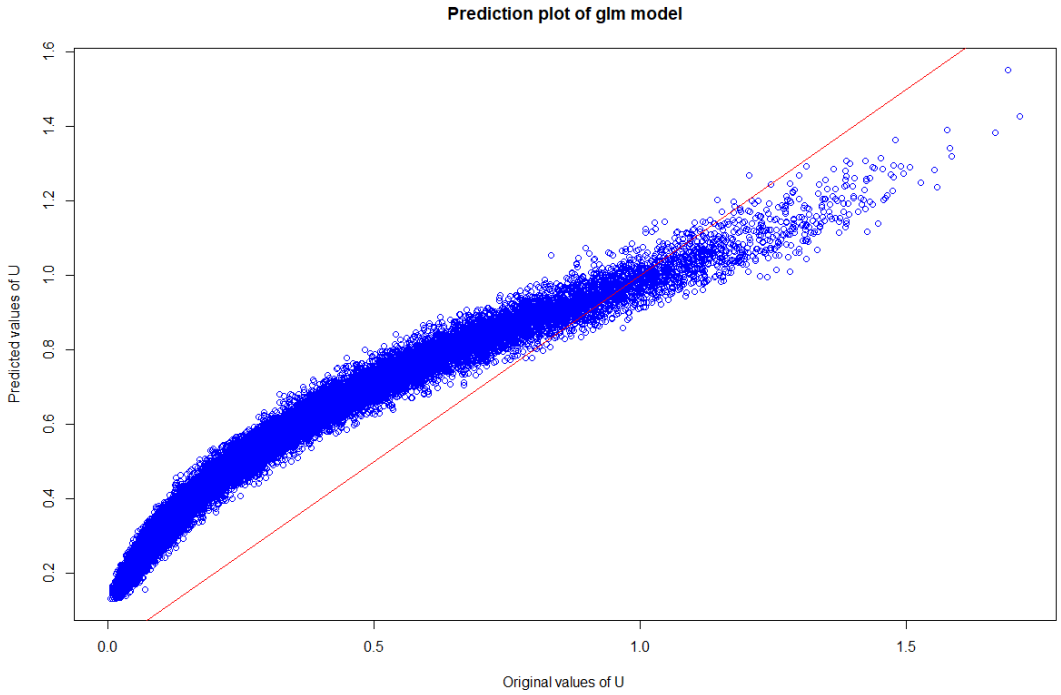


Figure 5.2: Prediction plot of glm model.

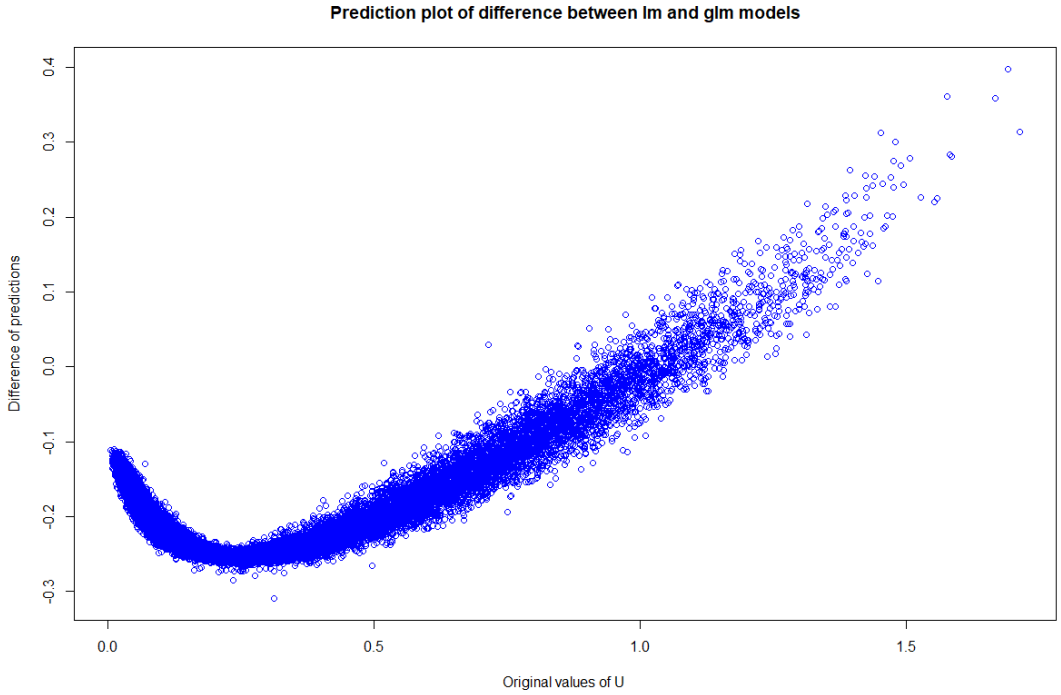


Figure 5.3: Plot of difference between predictions of lm and glm model vs. real uranium intensities.

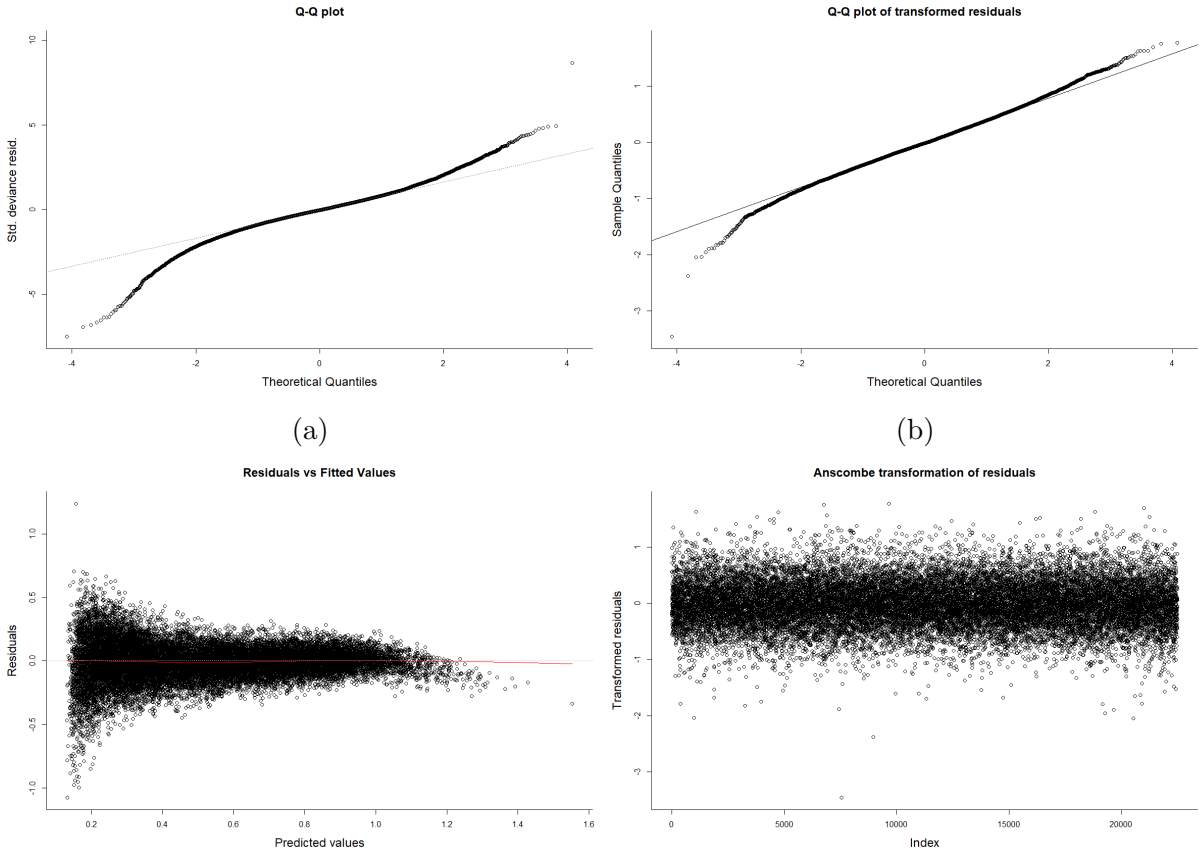
Summary of glm submodel:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.1e-01	7.1e-04	1.5e+02	0.0e+00
s	2.9e-02	3.4e-03	8.5e+00	1.4e-17
I(s^3)	4.6e-02	8.4e-03	5.5e+00	3.3e-08
z	1.1e-01	1.7e-03	6.4e+01	0.0e+00
I(z^2)	-8.1e-03	6.4e-04	-1.3e+01	1.2e-36
I(z^3)	5.2e-04	3.7e-05	1.4e+01	2.9e-45
a1	9.7e-03	3.2e-02	3.0e-01	7.7e-01
I($a1^3$)	2.5e+00	3.6e-01	6.9e+00	6.3e-12
a2	8.4e-02	8.9e-03	9.4e+00	7.2e-21
f1	-1.1e-01	1.0e-02	-1.1e+01	6.1e-27
I($f1^2$)	1.6e-01	2.9e-02	5.4e+00	5.2e-08
I($f1^3$)	-1.9e-02	2.0e-03	-9.5e+00	2.0e-21
f2	1.8e-02	3.9e-03	4.7e+00	2.5e-06
I($f2^2$)	-4.0e-03	2.5e-03	-1.6e+00	1.1e-01
I($f2^3$)	1.7e-03	1.2e-04	1.4e+01	4.7e-46
h1	8.5e-02	2.1e-02	4.1e+00	3.4e-05
I($h1^2$)	-2.1e-01	3.7e-02	-5.6e+00	2.2e-08
h2	8.7e-01	5.1e-02	1.7e+01	1.2e-65
I($h2^2$)	-3.5e+00	8.6e-01	-4.1e+00	4.7e-05
I($h2^3$)	9.0e+00	1.5e+00	5.9e+00	3.6e-09
nb	2.6e-01	6.9e-03	3.8e+01	9.7e-299
I(nb^2)	-3.3e-02	9.6e-03	-3.4e+00	6.3e-04
I(nb^3)	8.9e-03	1.8e-03	5.0e+00	5.7e-07
t1	1.2e-03	1.4e-02	9.0e-02	9.3e-01
I($t1^2$)	-1.5e-01	1.9e-02	-7.8e+00	8.6e-15
t2	4.8e-01	5.0e-02	9.6e+00	8.7e-22
I($t2^2$)	-4.8e+00	8.0e-01	-5.9e+00	3.1e-09
I($t2^3$)	1.3e+01	2.1e+00	6.1e+00	1.3e-09
s:z	-5.3e-02	3.8e-03	-1.4e+01	1.8e-44
s:f1	1.3e-01	2.6e-02	5.1e+00	3.2e-07
s:f2	-5.6e-02	9.8e-03	-5.7e+00	9.1e-09
s:h1	4.3e-01	6.0e-02	7.1e+00	9.2e-13
s:h2	-5.1e-01	1.3e-01	-4.1e+00	5.1e-05
z:f1	-3.0e-02	6.4e-03	-4.7e+00	3.1e-06
z:f2	2.0e-02	2.2e-03	9.0e+00	1.9e-19
z:h1	-2.6e-02	8.8e-03	-3.0e+00	2.6e-03
z:h2	-8.3e-02	3.2e-02	-2.6e+00	1.1e-02
z:t2	-1.2e-01	3.2e-02	-3.6e+00	3.4e-04
a1:a2	-7.1e-01	5.9e-02	-1.2e+01	6.1e-33
a1:h2	2.0e+00	3.6e-01	5.6e+00	2.0e-08
a2:f1	4.4e-02	1.9e-02	2.4e+00	1.9e-02
a2:f2	-1.4e-02	7.6e-03	-1.8e+00	6.7e-02
a2:h1	1.6e-01	5.6e-02	2.9e+00	4.1e-03
a2:h2	-3.2e-01	1.5e-01	-2.1e+00	3.9e-02

a2:nb	4.2e-02	1.1e-02	3.8e+00	1.5e-04
f1:f2	-6.6e-02	1.5e-02	-4.3e+00	1.7e-05
f1:h2	5.3e-01	1.9e-01	2.8e+00	5.0e-03
f1:t2	3.6e-01	1.3e-01	2.9e+00	4.3e-03
f2:h2	-1.7e-01	6.3e-02	-2.7e+00	6.6e-03
f2:nb	-3.3e-02	3.9e-03	-8.4e+00	3.2e-17
h1:t1	3.6e-01	4.3e-02	8.2e+00	2.1e-16
h1:t2	-5.9e-01	2.8e-01	-2.1e+00	3.2e-02

Null deviance	0.03699 on 21840.70 on 22494 degrees of freedom
Residual deviance	469.67 on 22443 degrees of freedom
AIC	-92770
Dispersion parameter	0.0205318

Table 5.2: Summary of glm submodel.



(a) (b)

(c) (d)

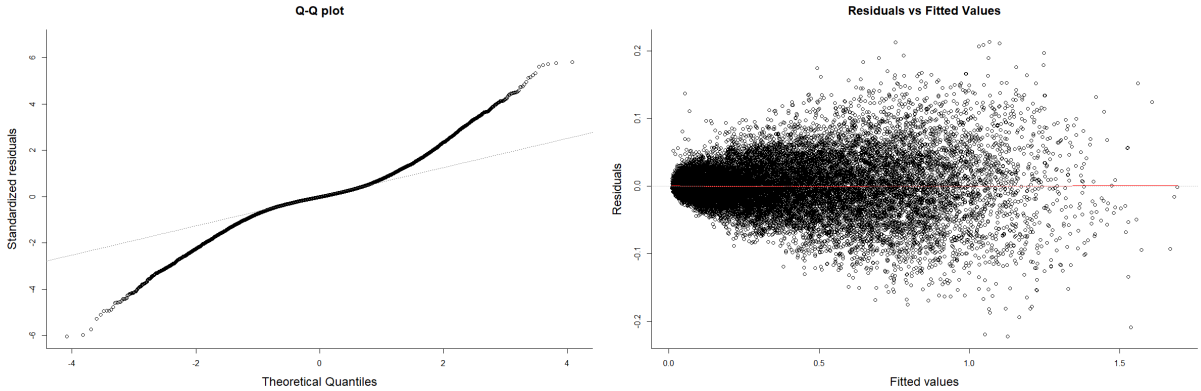
Figure 5.4: Quantile-quantile plot of residuals of glm model: (a) no transformation, (b) Anscombe transformation. Plot of residuals of glm model: (c) no transformation, (d) Anscombe transformation.

Suitable linear model without square root transformation of uranium has formula:

$$u \sim s + I(s^3) + z + I(z^2) + I(z^3) + a1 + I(a1^2) + I(a1^3) + a2 + I(a2^3) + f1 + f2 + I(f2^3) + h1 + h2 + I(h2^2) + nb + t1 + I(t1^2) + I(t1^3) + t2 + s : z + s : f1 + s : f2 + s : h1 + s : nb + z : a2 + z : f1 + z : f2 + z : nb + z : t1 + a1 : h2 + a1 : t1 + a1 : t2 + a2 : f1 + a2 : f2 + a2 : h1 + a2 : nb + a2 : t2 + f1 : h2 + f1 : nb + f1 : t1 + f1 : t2 + f2 : h1 + f2 : t2 + h1 : t1$$

Residual standard error	0.03699 on 22453 degrees of freedom
Multiple R-squared	0.9855
F-statistic	3.313e+04 on 46 and 22453 DF
p-value	$<$ 2.2e-16

Table 5.3: Summary of linear submodel without transformation.



(a) (b)

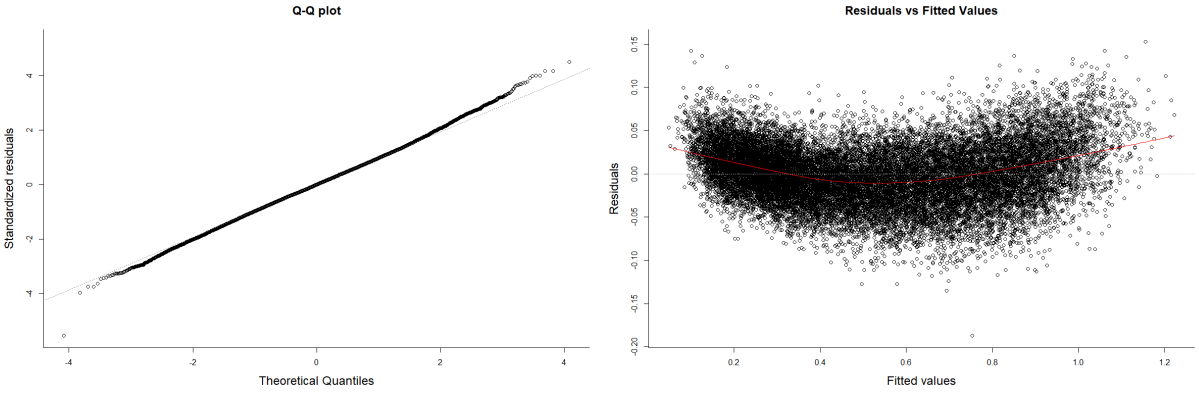
Figure 5.5: Q-Q plot and residual plot for linear submodel with no transformation.

Linear model of square root transformation of response and explanatory variables:

$sqrt(u) \sim sqrt(s) + sqrt(z) + sqrt(a1) + sqrt(a2) + sqrt(f1) + sqrt(f2) + sqrt(h1) + sqrt(h2) + sqrt(nb) + sqrt(t1) + sqrt(t2)$

Residual standard error	0.03425 on 22483 degrees of freedom
Multiple R-squared	0.9825
F-statistic	1.146e+05 on 11 and 22483 DF
p-value	$<$ 2.2e-16

Table 5.4: Summary of linear model with square root transformation of response and explanatory variables.



(a) (b)

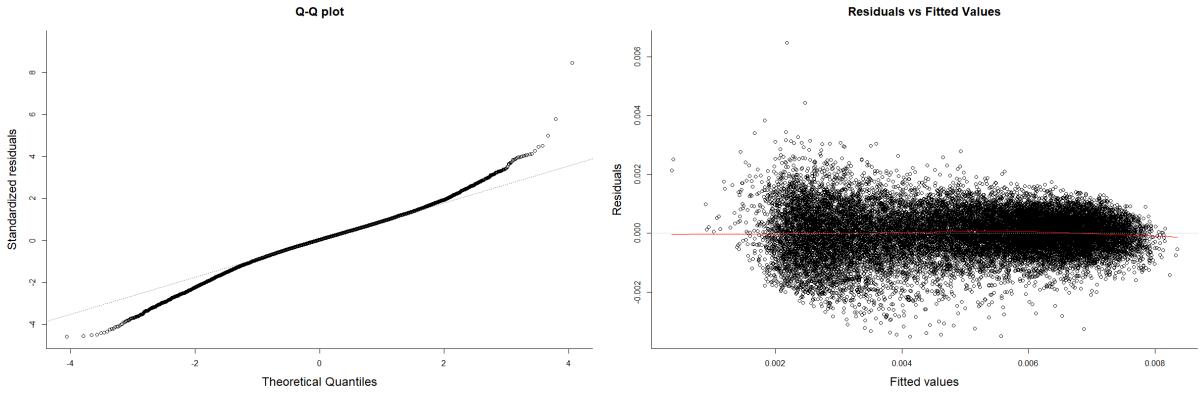
Figure 5.6: Q-Q plot and residual plot for linear model with square root transformation.

Linear model with standardised input data:

$$sqrt(u) \sim s + I(s^2) + I(s^3) + z + I(z^2) + a1 + I(a1^2) + I(a1^3) + a2 + I(a2^2) + f1 + I(f1^3) + f2 + I(f2^2) + I(f2^3) + h1 + I(h1^2) + I(h1^3) + h2 + I(h2^3) + nb + I(nb^2) + I(nb^3) + t1 + I(t1^2) + I(t1^3) + t2 + s : z + s : f1 + s : f2 + s : h2 + s : nb + s : t1 + s : t2 + z : h2 + z : nb + a1 : f1 + a1 : f2 + a1 : h2 + a1 : nb + a1 : t2 + a2 : f1 + a2 : f2 + a2 : h1 + a2 : h2 + a2 : nb + a2 : t2 + f1 : nb + f1 : t1 + f1 : t2 + f2 : t1 + f2 : t2 + h1 : nb + h1 : t1 + nb : t1 + nb : t2$$

Residual standard error	0.000769 on 19937 degrees of freedom
Multiple R-squared	0.8226
F-statistic	1650 on 56 and 19937 DF
p-value	$<$ 2.2e-16

Table 5.5: Summary of linear model with standardised input data.



(a) (b)

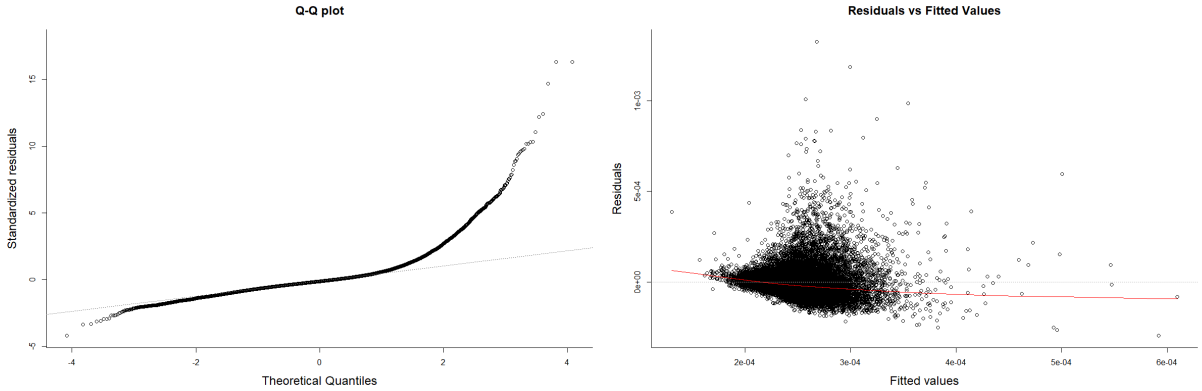
Figure 5.7: Q-Q plot and residual plot for linear model with standardised data.

Linear submodel with local standardisation of input data:

$$u \sim s + I(s^2) + I(s^3) + z + I(z^2) + I(z^3) + a1 + I(a1^2) + I(a1^3) + a2 + I(a2^2) + I(a2^3) + f1 + I(f1^2) + I(f1^3) + f2 + I(f2^2) + I(f2^3) + h1 + I(h1^2) + I(h1^3) + h2 + I(h2^2) + I(h2^3) + nb + I(nb^2) + I(nb^3) + t1 + I(t1^2) + I(t1^3) + t2 + I(t2^2) + I(t2^3) + s : z + s : a1 + s : a2 + s : f1 + s : f2 + s : h1 + s : h2 + s : nb + s : t1 + s : t2 + z : a1 + z : a2 + z : f1 + z : f2 + z : h1 + z : h2 + z : nb + z : t1 + z : t2 + a1 : a2 + a1 : f1 + a1 : f2 + a1 : h1 + a1 : h2 + a1 : nb + a1 : t1 + a1 : t2 + a2 : f1 + a2 : f2 + a2 : h1 + a2 : h2 + a2 : nb + a2 : t1 + a2 : t2 + f1 : f2 + f1 : h1 + f1 : h2 + f1 : nb + f1 : t1 + f1 : t2 + f2 : h1 + f2 : h2 + f2 : nb + f2 : t1 + f2 : t2 + h1 : h2 + h1 : nb + h1 : t1 + h1 : t2 + h2 : nb + h2 : t1 + h2 : t2 + nb : t1 + nb : t2 + t1 : t2$$

Residual standard error	8.131e-05 on 22411 degrees of freedom
Multiple R-squared	0.1134
F-statistic	32.58 on 88 and 22411 DF
p-value	< 2.2e-16

Table 5.6: Summary of linear submodel with local standardisation of input data.



(a) (b)

Figure 5.8: Q-Q plot and residual plot for linear submodel with local standardisation.