# TIMODAZ

Pavel Strnad

**Abstrakt:** Tento článek je zaměřen na analýzu dat z projektu TIMODAZ, který zkoumá vliv teploty na železobetonové ostění úložiště jaderného odpadu. V článku budou prezentovány především výsledky použitých metod při odstraňování odlehlých hodnot v pořízených datech a to, jakým způsobem byly chybějící hodnoty v datech doplněny. V závěru článku budou zhodnoceny výsledky a přidána doporučení.

**Klíčová slova:** TIMODAZ, datamining, odlehlé hodnoty, thresholding, interpolace, data inputting

**Abstract:** This article is focused on the data analysis of the project TIMODAZ, which examines the influence of temperature on the ferroconcrete concrete lining of the nuclear waste depot. The results of the measured data remote values elimination methods will be mainly presented in the article as well as how the missing values were supplemented. In the end of the article the results will be evaluated and recommendations given.

**Keywords:** TIMODAZ, datamining, outliers, thresholding, interpolation, data inpitting

**JEL Classification:** C8

## Cíle projektu

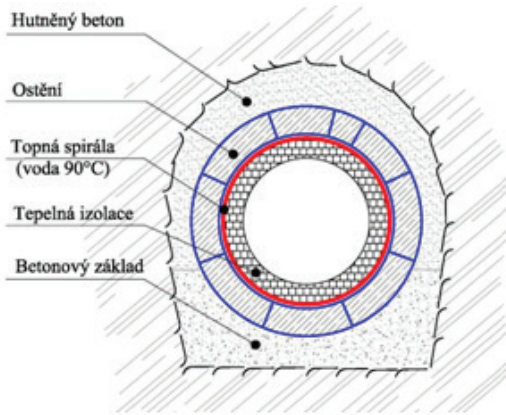
Cílem našeho projektu je důkladně zanalyzovat data z projektu TIMODAZ a ukázat na problémy, které se v datech objevují. Prozkoumat, zda v datech nejsou skryté závislosti. A na základě tohoto studia navrhnou metodiku pro pořizování, ukládání a zpracování dat u dlouhodobých experimentů.

## 1. Původ dat

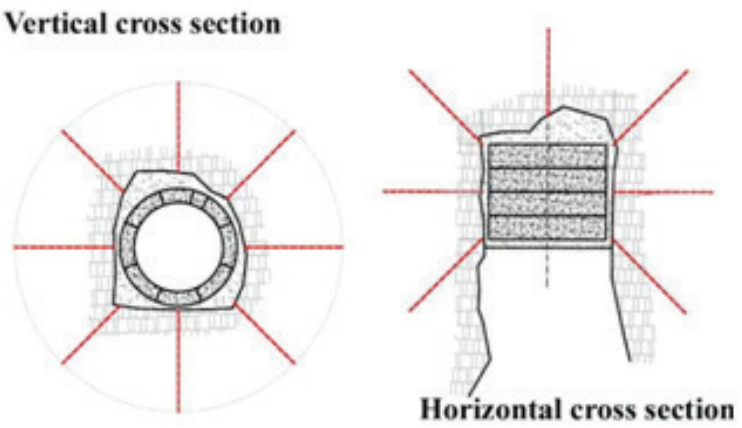
Data byla pořízena katedrou Centrum experimentální geotechniky při experimentálním pokusu v podzemním sile v laboratoři CEG (http://ceg.fsv.cvut.cz). Cílem výzkumu bylo ověřit, zda dlouhodobé tepelné zatížení nevyvolá v betonovém ostění úložného tunelu zatížení (napětí), které by vyčerpalo pevnostní charakteristiky betonu, a tím negativně ovlivnilo jeho stabilitu. [15]

*Obr. 1 uložení ve štole*

Fyzikální model byl vystavěn z prefabrikovaného ostění originálního belgického úložného tunelu dle projektu PRACLAY obr. 1 ve slepé štole, která byla vyražena technologií *„drill and blast"* (obr. 2) v tufitických horninách. Pro modelování extrémně nepříznivých okrajových podmínek byl prostor mezi montovaným ostěním a horninovým prostředím utěsněn výplňovým betonem. Betonová výplň mezi horninou a ostěním znemožňuje jednotlivým segmentům ostění i celým montovaným prstencům deformovat se vlivem tepelného zatížení směrem do horninového prostředí. Omezení (minimalizace) deformace vyvolá v betonových segmentech maximální napětí. Aby ostění zůstalo funkční (stabilní), nesmí tepelně indukované napětí překročit (vyčerpat) pevnostní charakteristiky materiálu ostění.[15]



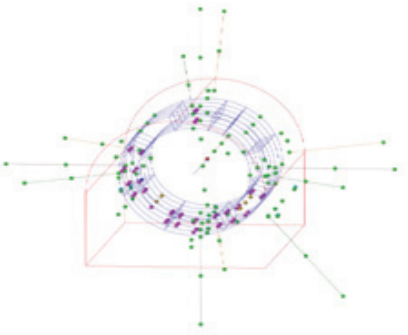

*Obr. 2 řez konstrukcí*




*Obr. 3 Orientace vrtů*

### 1.1. Výstavba modelu

Před zahájením vlastní výstavby byla v okolní hornině vytvořena síť vrtů pro instalaci teploměrů, které byly poté nainstalovány. Některá čidla (teploměry, tlakové čidla) byla instalována na spoj horniny a ostění tunelu. Po zhotovení samotného ostění tunelu byla instalována ostatní čidla a zařízení pro ohřev modelu. Obrázek instalovaných čidel naleznete níže.



*Obr. 4 uložení čidel*

### 1.2. Časový horizont pokusu

Pro „*in situ*“ experiment musel být splněn požadavek (SCK-CEN Belgie), že rozdíl teploty na vnitřní a vnější straně ostění nesměl překročit 30° C. Proto musel být „*in situ*“ model tepelně zatěžován postupně po jednotlivých krocích: 40° C (fáze 1), 60° C (fáze 2), 70° C (fáze 3), 80° C (fáze 4), 90° C (fáze 5). [15]

V budoucnosti, po vyhodnocení získaných dat z dosud aplikovaných zatěžovacích procedur, je u experimentu plánováno např. šokové cyklické zatěžování, čímž se extrémní zatěžovací podmínky ještě zintenzivní.

Celý dosavadní průběh experimentu je rozdělen na dvě testovací fáze (fáze 01, fáze 02) a pět zatěžovacích fází (přičemž fáze 5 nebyla dosud zahájena):

| | | | |
|---|---|---|---|
| 1 | Fáze 01 – měření během výstavby *„in situ" modelu* | | Od dubna 2008 |
| 2 | Fáze 02 – měření během testování topného systému | | Od 27. října 2008 |
| 3 | Fáze 1 – 1. zatěžovací stupeň | 10°C – 40°C | Od 30. října 2008 |
| 4 | Fáze 2 – 2. zatěžovací stupeň | 40°C – 60°C | Od 7. listopadu 2008 |
| 5 | Fáze 3 – 3. zatěžovací stupeň | 60°C – 70°C | 5. února 2009 |
| 6 | Fáze 4 – 4. zatěžovací stupeň | 70°C – 80°C | duben 2009 - květen 2009, červen 2009 – říjen 2009 |
| 7 | Fáze 5 – 5. zatěžovací stupeň | 80°C – 90°C | |

*Tabulka 1 Fáze experimentu*

## 1.3.Prvotní analýza dat

Při prvotní analýze dat jsme provedli kategorizaci čidel dle meta dat a změřili počty chybějících hodnot jimi pořízené. Výsledky prvotní analýzy je možné vidět v tabulce 1. Jak je vidět níže, procentuální vyjádření chybějících hodnot je 23,13 % což není velké číslo. Bohužel měly některé atributy výrazně větší počty chybějících hodnot. V procentuálním vyjádření se jednalo o 20-30% úspěšnost pořízení dat na konkrétním čidle.

| | | | |
|---|---|---|---|
| Čidla celkem | 284 | 40 - reg. teploty | 2 |
| Počet typů | 11 | 55 - teploměr | 57 |
| Z toho | | 60 - tenzometr | 12 |
| 10 - teploměr | 124 | 65 - odpor | 2 |
| 20 - posun | 8 | 70 - tl. buňka | 4 |
| 25 - zdroj | 2 | Celkový počet dat | 45 529 770 |
| 30 - kotel | 4 | Z toho chybějících | 10 534 788 = 23,13% |
| 35 - el. | 9 | | |

Tabulka 2 přehled čidel

## 2. Architektura řešení DWH

Pro řešení byly použity následující nástroje. RapidMiner [9] jako nástroj pro spouštění jednotlivých částí procesů. Dále byla použita databáze MySQL [8] jako úložiště dat. Pro řešení byla použita vícevrstvá architektura datového skladu. [1] Důvodem vytvoření vícevrstvé datové struktury bylo maximální zjednodušení procesu načítání a čištění dat. Dalším důležitým aspektem řešení je, že uživatel má možnost sledovat tři druhy dat.

1) Data pořízená primárním systémem,

2) Data vyčištěná v netransformované podobě

3) Data přepočtena do skutečných hodnot pomocí pracovních diagramů sledovaných materiálu.

Výsledkem je minimální latence při dotazování na jednotlivé datové vrstvy. A to z toho důvodu, že veškeré výpočty probíhají rutinně v noci. Uživatel se dívá pouze na výsledky a nezatěžuje DB server složitými on-line výpočty. Jak vyplývá z předcházející věty, data nejsou načítaná on-line, ale s 24 hodinovým zpožděním.

### 2.1. RapidMiner

Z vrstvy L0 je načtena tabulka L0_mereni. Tato tabulka obsahuje data jednotlivých čidel od počátku pokusu do posledního load-u dat. V RM byla provedena další analýza. V prvním kroku byla data pouze načtena a zkoumána za pomocí grafů. Na ose X byla vynesena časová osa a na osu Y jednotlivé naměřené hodnoty konkrétního čidla. U většiny čidel bylo zjištěno, že naměřená data obsahují velké množství nekorektních měření (odlehlých hodnot). Důvodem vzniku odlehlých hodnot je vada čidel během pokusu. Jak bylo sděleno pracovníky z katedry CEG. Tyto odlehlé hodnoty je nutné v naší úloze odstranit pro následné doplňování hodnot.

V případě, že data budou obsahovat velký počet odlehlých hodnot, tak následné doplnění muže mít nežádoucí účinek. A to ten, že by se mohly doplnit hodnoty, které jsou také odlehlé.

## 3. Odlehlé hodnoty

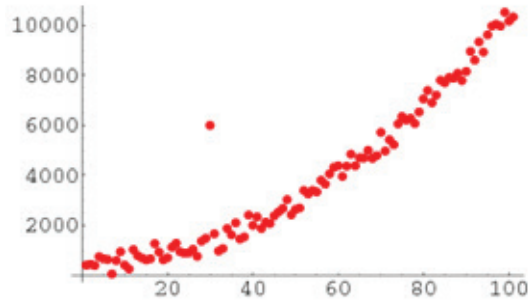
Co je to odlehlá hodnota. K tomu to pojmu je možné nalézt velké množství definicí. Nejvíce vystihující jsou následující:

- Odlehlá hodnota je takové pozorování, které se od ostatních liší natolik, že se tváří, jako by bylo naměřeno jiným zařízením. [7]
- Odlehlá hodnota je takové pozorování (nebo skupina pozorování), které se zdá být
- v rozporu se zbytkem naměřených pozorování. [10]

Z jaké příčiny vznikají odlehlé hodnoty? Příčin je hned několik:

1) špatně kalibrované zařízení pro požívání údajů
2) špatně zvolený postup měření
3) lidský faktor
4) vnější vlivy prostředí

Jak tyto odlehlé hodnoty odhalovat. V případě pořizování dat, která nemají velkou dynamiku a výraznou korelaci, je možné data jednoduše rozpoznat z pohledu na graf se dvěma proměnnými. Jak je možné vidět na obrázku níže.




*Obr. 5 Odlehlé hodnoty*

Tato situace je pouze ideální případ, který se při měření dlouhých časových řad vyskytuje velmi zřídka. Většinou se data pořídí s velkým počtem odlehlých hodnot, které nám při vyhodnocování výrazně zkreslují vlastnosti měřeného procesu. Proto je velmi důležité tyto hodnoty nalézt a odstranit.

## 4. Metody použitelné pro odstranění odlehlých hodnot

V následujícím odstavci naleznete nejpoužívanější metody pro odstraňování odlehlých hodnot v časové řadě.

### 4.1. K-means

K-means je metoda shlukové analýzy, jejímž cílem je rozdělit n pozorování do K-tříd, přičemž každé pozorování patří do konkrétní třídy.

Základní parametr pro tuto metodu je K. Tento parametr udává, do kolika tříd budeme chtít data zatřiďovat. Základní algoritmus funguje tak, že se náhodně vygeneruje tolik objektů (středů clusteru) dle počtu K. Následně se prohledává prostor v okolí a body, které se nejvíce podobají konkrétnímu objektu, jsou do této množiny zařazeny.

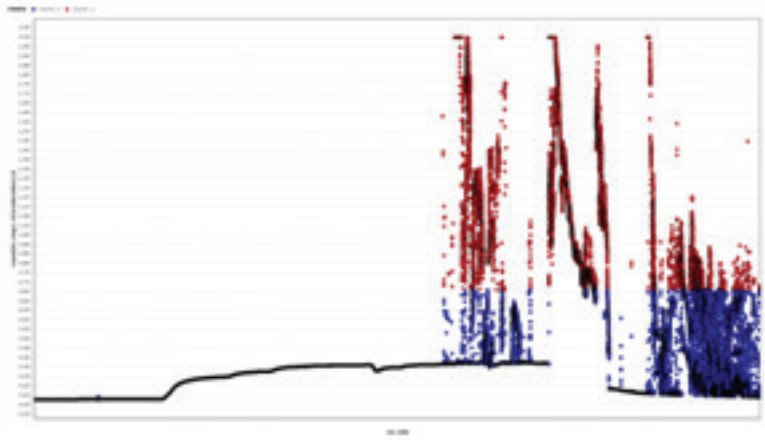
Tento postup probíhá iterativně, dokud nezačne kriteriální funkce konvergovat. Nejčastěji se používá kriteriální funkce nejmenších čtverců.

$$E = \sum_{i=1}^{k} \prod_{x \in C_i} |x - m_i|^2$$

Výsledkem této metody je rozdělování datového prostoru do Voronoi diagramu, ve kterém je možné vidět rozdělení dat do tříd. V případě odstraňování odlehlých hodnot se snažíme separovat odlehlé hodnoty od ostatních dat.

U této metody bohužel dopředu nevíme, na kolik tříd data rozdělit, a následně vybereme potřebou třídu obsahující validní hodnoty. Tato metoda se nejčastěji používá k rozpoznávání vzorů v datech, zpracování signálu, web miningu a odstraňování odlehlých hodnot [12].

Při využití této metody na odstranění odlehlých hodnot v časové řadě se nám nepovedlo vhodně implementovat tuto metodu do procesu čištění dat. Výsledkem sice bylo zatřídění do k tříd, ale v globálním pohledu na data se nepovedlo dosáhnout uspokojivých výsledků. Použití metody v globálním pohledu na data, není použitelná. V případě použití malého intervalu hodnot byly výsledky lepší, ale stále neuspokojivé.



*Obr. 6 Použití metody K-means*

Závěr: tato metoda na naše data není vhodná, jelikož se nejedná o klasifikovatelná data do jednotlivých tříd, jak ukazuje obrázek 6.

## 4.2. Regresní modely

Lineární regrese s jednou vysvětlující proměnnou

Vycházíme ze vzorce:

$$Y = f(x) + e$$

kde e je náhodná chyba vzniklá jako chyba měření nebo působením jiných náhodných vlivů. Veličina Y je zřejmě náhodná, neboť vzniká jako součet nenáhodné funkce f(x) a náhodné chyby e.

Hledáme regresní funkci f, známe-li n dvojic (x1,y1), … , (xn,yn), kde xi je hodnota nezávislé, vysvětlující proměnné Xi a yi hodnota odpovídající závislé, vysvětlované proměnné Yi přičemž předpokládáme, že

$$Yi = f(xi) + ei,\ i = 1, \ldots, n.$$

Víme-li předem, že regresní funkce f(x) má tvar

$$f(x) = b0 + b1x,$$

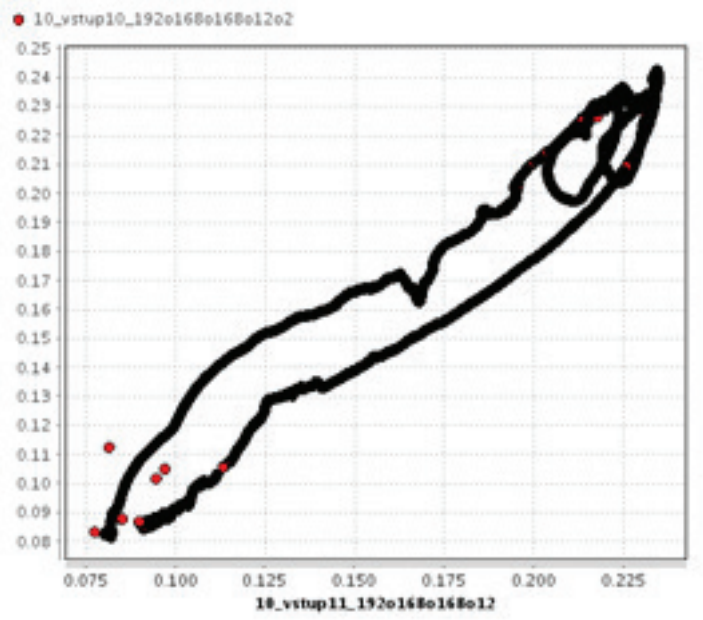
pak mluvíme o lineární regresi s jednou vysvětlující proměnnou, případně o jednoduché lineární regresi.

Ostatní regresní modely jsou dost podobné a vycházejí z lineární regrese. Jde jen o to odhadnout, který regresní model se má na daná data použít.

Regresní modely:

- lineární regrese
- kvadratická regrese
- exponenciální regrese
- polynomická regrese

Na základně prvotní analýzy jsme zjistili, že vztahy mezi jednotlivými čidly nelze triviálně popsat, jak je vidět na grafu č. xxx. Bylo by nutné data rozdělit do jednotlivých intervalů a v každém intervalu vytvořit regresní model. Problém je v tom, jak automaticky určit vhodný interval. Na základě těchto skutečností použijeme nejdříve jednodušší metodu, u které nebude tolik nejistých výsledků.




*Obr. 7 Závislost dvou teplotních čidel*

### 4.3. Dynamika časové řady

Tato metoda je založena na sledování změny (diference) v předchozích hodnotách vůči aktuálně sledovanému měření.

Předpokládejme časovou řadu yt, t = 1, ..., T. Nejjednodušší mírou dynamiky je absolutní přírůstek (první diference), který lze zapsat jako

$$\Delta yt = yt - yt\text{-}1,\ t = 2, ..., T$$

Tato charakteristika vyjadřuje změnu hodnoty v čase t proti času t -1. Často se používá také průměrný absolutní přírůstek

$$\bar{\Delta} = \frac{(y_2 - y_1) + (y_3 - y_2) + \ldots + (y_T - y_{T-1})}{T-1} = \frac{\sum_{t=2}^{T} \Delta y_t}{T-1} = \frac{y_T - y_1}{T-1}.$$

Ten se pak používá, jako maximální míra diference při vyhodnocování hodnoty absolutního přírůstku. Hodnota absolutního přírůstku nemusí nutné používat yt-1 ale hodnoty t-x, kde jako x můžeme zvolit libovolnou kladnou hodnotu z oboru reálných čísel. V RM pro tuto metodu byly použity následující komponenty. (Retrive, Set Role, Series: Differentiate (Series), Filter Examples).

Tato metoda při testování vrací zatím nejlepší výsledky. Je zde ale jisté omezení v případě náhodných odlehlých hodnot. Metoda odstraní potřebné nepřesnosti, ale v situaci, kdy chybovost má jistý průběh (OBR xxx), odstraní pouze část chyb. Proto je nutné nalézt další metodu, která následně odstraní odlehlé hodnoty, nebo zcela nahradí doposud preferovanou metodu.

### 4.4. Moving Avarage

Jedná se o velmi známou techniku klouzavého průměru. Metoda funguje na výpočtu průměrování plovoucího okénka o velikosti w, v kterém se vypočítavá průměrná hodnota konkrétního okénka. Metod výpočtu je několik. A liší se požadovaným typem průměru. Po výpočet X' je původní X nahrazeno X'.

1) počátek intervalu

$$X' = \left( x + \sum_{i=1}^{i=w} \frac{\lfloor x_i \rfloor}{w} \right]$$

2) střed intervalu

$$X' = \frac{\sum_{i=-\frac{w}{2}}^{i=-1} x_i + x + \sum_{i=1}^{i=\frac{w}{2}} x_i}{w}$$

3) konec intervalu

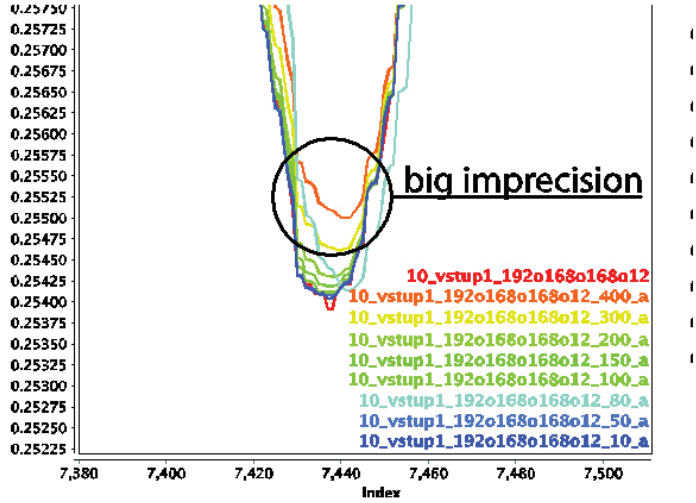
$$X' = \left( x + \sum_{i=w}^{i=1} \frac{[x_i)}{w} \right]$$

Další hledisko výpočtu je výpočet vzdálenosti prvků v prostoru. Pro výpočet je možné použit řadu metrik. Pro příklad euklidovská vzdálenost (Euclidean distances), trojúhelníková vzdálenost (triangle distances), čtvercová (square distances) a mnoho dalších. Pro redukci odlehlých hodnot byla použita komponenta Moving Average. Tato metoda vytváří průměrnou hodnotu z definované šířky plovoucího okénka. Tato metoda má několik vstupních parametrů., které můžete vidět na obrázku xxx.

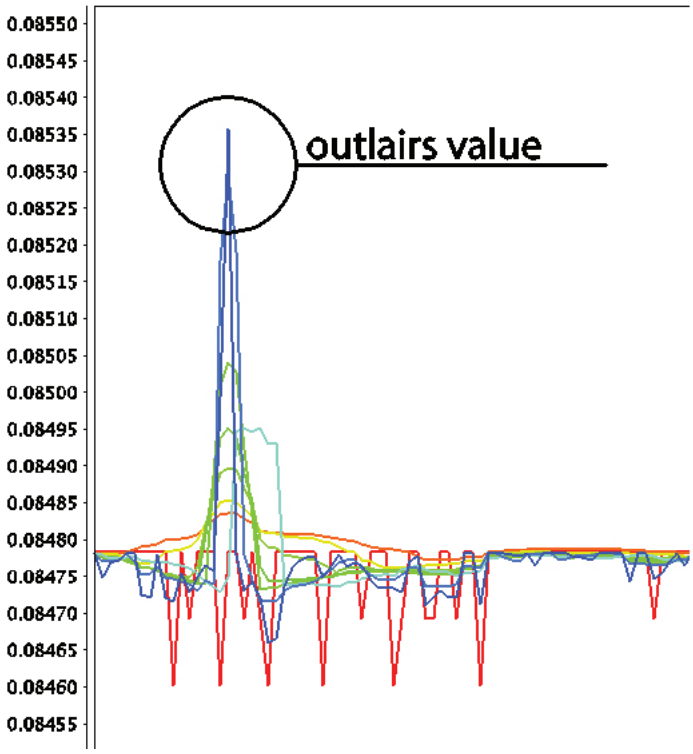
Pro pokus byly vytvořeny dvě skupiny dat. Kompletní datový soubor (v grafech označen jako _a) a vzorek reprezentovaný 5000 hodnotami (v grafech označen jako _s). Na těchto datech byla použita metoda Moving Average s rozdílnou hodnotou velikosti okénka. Hodnoty byly nastaveny na 10, 50, 80, 100, 150, 200, 300, 400 pro obě skupiny dat.

Výsledky této metody je možné vidět na obrázcích 7, 8, 9, 10. Je vidět, že průměrování odlehlých hodnot s malou odchylkou funguje velmi dobře, ale v případě, kdy data mají extrémní hodnotu odchylky, metoda selhává a v okolí extrému jsou data zcela znehodnocena a nabývají nerelevantních hodnot. Pro ilustraci je zde obrázek 8. Jak je z obrázku zřejmé, při vyšší šířce okénka dochází k lepšímu vyhlazení grafu. V částech, kde nejsou velké změny, jsou data dokonale vyrovnána, ale v místech, kde je radikální změna, se použité metody značně liší. Při použití menší šířky okénka nejsou data tak dobře vyhlazena, ale na druhou stranu dobře reagují na velké změny v průběhu. Tuto metodu je vhodné použít v kombinaci

s dalšími metodami, jak se dočtete níže.




*Obr. 8 chyba při velkém hodnotě okénka*




*Obr. 9 Odlehlé hodnoty*


*Obr. 10 Chybějící hodnoty*

### 4.5. Prahování (TrashHolding)

Tato metoda je založena na filtrování vstupních hodnot za pomoci vybraných atributů a nastavením hodnoty, která bude rozlišovat, zda se jedná o data validní či ne.

$$f(c) = \begin{cases} A & \text{pokud } c < práh \\ B & \text{pokud } c \geq práh \end{cases}$$

## 5. Doplňování chybějících hodnot

Pro doplňování chybějících hodnot se používá celá řada metod. V přehledu bude stručně vysvětleno, jakým způsobem funguje a proč tato metoda není vhodná pro doplňování chybějících hodnot v našich datech. [17]

**Zero imput** - náhrada chybějící hodnoty hodnotou = 0. Tato metoda je velmi jednoduchá a v některých případech může splnit požadavky. Bohužel tato metoda nemohla být použita z toho důvodu, že v měřených datech se vyskytují hodnoty 0, a také by tyto hodnoty ve většině případu vytvořily nové odlehlé hodnoty.

**Max imput -** vybere maximální hodnotu v datovém souboru a doplní jí na všechny chybějící hodnoty. Tato metoda při realizaci na naše data byla nevhodná, jelikož vygenerovala další odlehlé hodnoty

**Min imput** - obracená metoda Max imput

**Average imput** - tato metoda je založena na výpočtu průměrné hodnoty. Tato metoda také není zcela vhodná pro doplňování do dat v našich datech.

**Value imput -** vkládání konkrétně zadané hodnoty.

**Median imput** - metoda založena na základní statistické metodě. Výsledkem je doplnění nejčastěji vyskytující se hodnoty v našich datech.

**Next value imput** - doplní hodnotu svého předchůdce od pozice, na které leží. Tato metoda je v datech s velmi malou změnou průběhu dobře použitelná

**Preview value imput** - doplní hodnotu svého následníka od pozice, na které leží. Tato metoda je v datech s velmi malou změnou průběhu dobře použitelná

**Randon generator** - generuje náhodné hodnoty na stanoveném intervalu. V případě našich dat by se jednalo o vygenerování odlehlé hodnoty

**Linear interpolation imput -** tato metoda je založena na dopočítávání hodnot mezi dvěma hodnotami. V našem případě se tato metoda jeví jako velice efektivní. Za jistých předpokladů:

- v datech nejsou výrazně změny průběhů, které by díky lineární interpolaci nebyly zohledněny
- dokonalé odstranění odlehlých hodnot

$$\frac{y - y_0}{y_1 - y_0} = \frac{x - x_0}{x_1 - x_0}$$

**Regrassion model [11]** - druhy regresních modelu naleznete v kapitole 5.4

## 6. Postup čištění dat TIMODAZ

Pro data z projektu TIMADAZ jsme zvolili následující postup čištění dat. V prvním kroku jsou odstraněny odlehlé hodnoty a následně do vyčištěných dat jsou chybějící data doplněna

### 6.1. Postup č. 1

Tento postup jsme pracovně pojmenovali jako MAT (Moving Average, Thresholding). Postup čištění dat je založen na metodě klouzavého průměru spojené s metodou prahování. [16]

**Postup:**

1) pomocí metody Moving Average [odkaz] vytvoříme nový dataset, který je zprůměrován svým okolím.

2) vytvoříme nový atribut delta, který vypočítá absolutní odchylku nově vytvořeného datasetu od původního.

$$\Delta d = |x' - x| \quad (1)$$

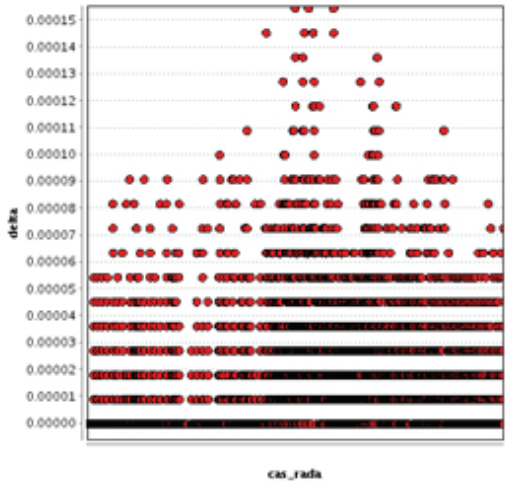
x - původní data
x' - data vytvořena pomocí "*moving average*"

3) V tomto kroku provádíme prahování a body které nesplňují podmínku (2) jsou z datasetu vyloučeny.

$$\Delta d < h \quad (2)$$

h - prahová hodnota, h = 0,00005

Velikost hodnoty y byla nastavena na základě grafu Δ obr. 11, na kterém je vidět, že frekvence výskytu větších odchylek je menší než u chyb s menší hodnotou odchylky, a proto stanovená hranice je dostatečná.




*Obr. 11 Odlehlé hodnoty delta*

4) V posledním kroku doplňujeme chybějící hodnoty pomocí lineární interpolace.

## 6.2. Postup č. 2

Tento postup jsme pracovně pojmenovali **TMAT** (Thresholding, Moving Average, Thresholding). V této metodě používáme kombinaci metod prahování, výpočet delty z předchůdců a výpočet delty vůči datům.

Výpočet delty z předchůdců je počítán tak, že je porovnána hodnota svého y-tého předchůdce. Hodnoty y je možné určit. V našem pokusu jsme zvolili hodnotu 6, aby se porovnávání blížilo hodinovým průměrům (6x10minut).

$$\Delta p = |xi - (xi-y)| \quad (3)$$

y - počet předchůdců

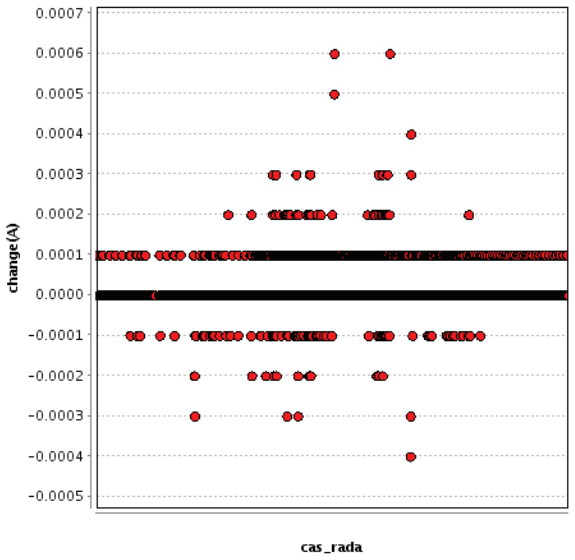
**Postup:**

1) Výpočet delty z předchůdců

2) V tomto kroku provádíme prahování a body, které nesplňují podmínku (4), jsou z datasetu vyloučeny.

$$\Delta p < h \quad (4) \text{ , } h = 0{,}003$$

Velikost hodnoty byla nastavena na základě grafu Δ obr. 12




*Obr. 12 delty z metody TMAT*

3) Po hrubém vyčištění v kroku 2 použijeme metodu Moving Average, která data vyhladí.

4) výpočet Δd (1)

5) prahování

$$\Delta d < h\ (2),\ h = 0{,}00005$$

6) doplnění hodnot pomocí lineární interpolace

### 6.3. Postup č. 3

Tento postup jsme pracovně pojmenovali T (Thresholding).

**Postup**

1) výpočet Δp

2) prahování Δp

$$\Delta p < h\ (4),\ h = 0{,}003$$

3) doplnění hodnot pomocí lineární interpolace

### 6.4. Postup č. 4

Tento postup jsme pracovně pojmenovali **MATMA** (Moving Average Thresholding Moving Average).

Postup

1) pomocí metody Moving Average vytvoříme nový dataset

2) výpočet $\Delta d$

3) prahování

$$\Delta d < 0{,}00005 \quad (2)$$

4) pomocí metody Moving Average vytvoříme nový dataset, který dále považujme za požadovaná data

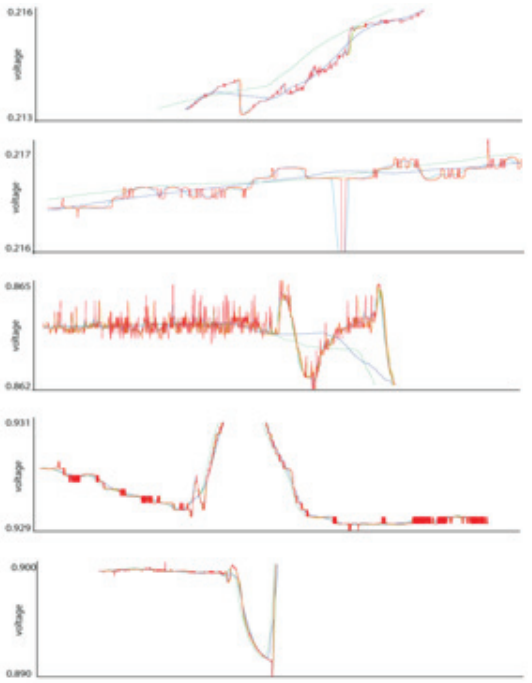
5) doplnění hodnot pomocí lineární interpolace

### 6.5. Hodinové průměry

Agregací dat spočteme průměry za jednotlivé hodiny.

### 6.6. Denní průměry

Agregací dat spočteme průměry za jednotlivé dny.



*Obr. 13 Agregace dat*

## 6.7. Vyhodnocení postupů

Hodnocení postupu z hlediska chybějících hodnot je vidět v tabulce 3. Každá metoda odstraní jiný počet odlehlých hodnot. U neagregačních metod se odstraní zanedbatelné množství dat proti celkovému počtu 159195 dat. Naopak agregační hodnoty výrazně sníží počet dat na 26533, ale zároveň vnesou chybu do výsledných dat. Která z kategorie je vhodnější není možné jednoznačně určit. Záležet hlavně bude na osobě, která data bude následně používat.

| Atribute | missing value | Filtred value | Filtr missing % |
|---|---|---|---|
| cas_rada | 0 | 0 | 0.00% |
| Originál | 11883 | 0 | 0.00% |
| MAT | 20967 | 9084 | 5.71% |
| TMAT | 20594 | 8711 | 5.47% |
| T | 12673 | 790 | 0.50% |
| avg(day) | 158153 | 146270 | 91.88% |
| avg(hour) | 134596 | 122713 | 77.08% |
| MATMA | 20967 | 9084 | 5.71% |

*Tabulka 3 Chyby v datech*

Vyhodnocení metod z hlediska čistoty dat

| atribut | miss | relativ miss % |
|---|---|---|
| cas_rada | 0 | 0.00% |
| avgMAT(day) | 2091 | 7.88% |
| avgT(day) | 1997 | 7.53% |
| avg MATMA (day) | 2091 | 7.88% |
| avg TMAT (day) | 2033 | 7.66% |

*Tabulka 4 Relativní chyby v datech*

## Zjištění

Od prvotního předpokladu, že hodnotu delta můžeme nastavit na pevno, musíme odstoupit, protože každý typ čidla má diference v jiných řádech jako ostatní čidla (teplotní čidla v řádech desítitisícin oproti tlakovým čidlům v řádech stovek). Z tohoto důvodu je potřeba deltu odchylek chytře spočítat:

- buď ji spočítat pro jednotlivé typy čidel
- pro každý dataset daného čidla ji počítat v navrženém procesu pomocí kvantilů na zvolené hodnotě alfa

Na základě zkoumání diferencí odchylek při teplotních čidlech, kde byla delta nastavena na pevnou hodnotu, jsme odfiltrovali přibližně 10 % dat, která uvažujeme jako odlehlé hodnoty. Na tomto základě budeme uvažovat, že vyloučíme horní 10% alfa kvantil z hodnot delta. Jak je vidět z tabulky tab xxxzzz ve sloupci **relative miss %.**

### 6.8. Doplňování chybějících hodnot

Dalším krokem čištění je doplnění chybějících hodnot. Pro naši problematiku jsme použili následující postup doplňování chybějících hodnot

1) Lineární Interpolace
2) Doplnění prev value
3) Doplnění next value
4) V případě, že sloupec obsahuje pouze NULL, je doplněna hodnota 9999. Z důvodu následné logiky výpočtu do vrstvy L2

Pro tyto metody je brán předpoklad, že se měřené veličiny skokově nemění a jejich průběh je buď jen rostoucí, nebo klesající v intervalu jednoho týdne.

Po doplnění je soubor uložen do textového souboru a následně importován do DB. Toto řešení se provedlo z důvodu rychlosti exportu dat z RM a načtení do MySQL DB. Kompletní schéma procesu můžete vidět na obrázku

## 7. Vize do budoucna

Pomocí data mining metod se pokusit ověřit, zda se teplo v tunelu a jeho okolí šíří rovnoměrně všemi směry nebo se šíří ve smyslu od zemského jádra -'nahoru'. Na základě zjištěných skutečností navrhnou metodiku, která by univerzálně řešila problém ukládání a struktury dat v dlouhodobých pokusech a maximálně zjednodušila a urychlila následný proces zpracování dat do požadovaného výstupu.

## Závěr

V současné době je vytvořen datový sklad, který data transformuje do požadovaných domén. Výsledkem je tabulka L2_mereni která obsahuje 45,5 mil řádku. A skládá se z následujících atributů: ID,PERIODA,DATUM,CIDLO,HODNOTA. Na těchto datech byly provedeny základní testy čistoty pořízených dat, a jak bylo zmiňováno v předchozích kapitolách, DWH je připraven do fáze vývojového testovaní na katedře K128.

## Literatura


[1] Bob Griesemer , Oracle Warehouse Builder 11g R2: Getting Started 2011,Publisher: Pa,ckt Pub,lishi,ng 2011, 424 pp.

[2] Jingke Xi, Outlier Detection Algorithms in Data Mining, Intelligent Information Technology Application, 2008. IITA '08. Second International Symposium on Date of Conference: 20-22 Dec. 2008, Page(s): 94 - 97

[3] Svetlana Cherednichenko, Masters thesis, 2005
University of Joensuu Faculty of Science / Department of Computer Science, Outlier Detection in Clustering

[4] Josef Arlt, Markéta Arltová, Eva Rublíková , Analýza ekonomických řad s příklady, http://nb.vse.cz/~arltova/vyuka/crsbir02.pdf

[5] Moravec, Miroslav, Hledání parametrů modelů dynamických, http://hdl.handle.net/10195/39747

[6] Dantong Yu ,Finding Outliers in Very Large Datasets, Knowledge and Information Systems (2002) 4: 387-412

[7] D. M. Hawkins, "*Identification of Outliers*". Chapman and Hall, London, 1980.

[8] B.Schwartz, P. Zaitsev, V. Tkachenko, J. D. Zawodny, A. Lentz, D. J. Balling, High Performance MySQL, 304 pp.
Publisher: O'Reilly Media, Released: June 2008

[9] http://rapid-i.com/content/view/36/210/lang,en/

[10] V. Barnett and T. Lewis,Outliers in statistical data,1994, 3rd edition, (John Wiley & Sons, Chichester), 584 pp.

[11] Yang C. Yuan, Multiple Imputation for Missing Data: Concepts and New Development, SAS Institute Inc., Rockville, MD

[12] Yinghua Zhou, Hong Yu, Xuemei Cai, "*A Novel k-Means Algorithm for Clustering and Outlier Detection*," fitme, pp.476-480, 2009 Second International Conference on Future Information Technology and Management Engineering, 2009

[13] Jarušková, D., skriptum Pravděpodobnost a matematická statistika 12, ČVUT 2002, s.100-103

[14] http://is.muni.cz/el/1433/ podzim2006/PA034/ 09_SVM.pdf , dne 3.20.2012

[15] Jaroslav Pacovský , Radek Vašíček, Markéta Levorová, Výzkum bezpečné funkce ostění úložného tunelu hlubinného úložiště dlouhodobě zatíženého teplotou, 2009

[16] http://www.pvv.org/~perchrh/ papers/datasyn/paper2/report.pdf

[17] Andrew Gelman , Data Analysis Using Regression and Multilevel/Hierarchical Models, Missing-data imputation, Cambridge University Press 2006 pp. 529-544



*Ing. Pavel Strnad, FSV-ČVUT v Praze*

*Ing. Michal Višňovský, FSV-ČVUT v Praze*