

UKÁZKA STUDENTSKÉ PRÁCE V RÁMCI SOČ (METODA SAVITZKY-GOLAY)

Radek Hampl¹

Abstrakt:

V příspěvku je popsán způsob práce se studentem VOŠ a SPŠ ve Varnsdorfu, se kterým jsem se jako konzultant zúčastnil celostátní přehlídky SOČ. Tématem práce byla metoda Savitzky-Golay pro zpracování statistického souboru dat a oborem, ve kterém soutěžil, byla matematika. Jedná se tedy o ukázkou práce s talentovanými studenty.

1. Úvod

Podpora talentovaných studentů by měla být jednou z priorit vzdělávacích institucí všech typů, od základních škol, přes školy střední až po školy vysoké. Bohužel se na tuto velmi důležitou aktivitu občas zapomíná a tak není na škodu si připomenout, že mezi našimi studenty jsou i velmi talentovaní jedinci, kteří jsou schopni i na středních školách skutečně vysoce odborné práce daleko přesahující rámec učiva.

Jednoho svého studenta jsem tedy před dvěma lety (školní rok 2006 / 2007) oslovil s tím, že bylo zajímavé se zúčastnit soutěže SOČ (Středoškolská odborná činnost). Tím studentem byl Ondřej Tůma, který v té době na naší škole studoval obor Elektronické počítačové systémy (nyní je již po maturitní zkoušce a studuje na TU v Liberci). Vzhledem k tomu, že učím matematiku, obor byl jasný. Již dříve jsem se setkal s metodou Savitzky-Golay pro zpracování statistických dat. Tato metoda je starší nicméně se i nyní s úspěchem používá při zpracování difrakčních měření na polykrystallických materiálech ve Fyzikálním ústavu Akademie věd ČR, v. v. i.

2. Krátce o metodě Savitzky-Golay

Tuto metodu v roce 1964 představili právě pánové Savitzky a Golay. Jedná se o upravenou metodu MNČ. V době vzniku byla výpočetní technika v plenkách a i ty nejvýkonnější počítače měly nesrovnatelně nižší výkon než dnešní, byť jen běžné, stolní PC. Tehdejší počítače také disponovaly s velmi malou operační pamětí.

To znamenalo ale zásadní omezení pro ukládání větších objemů dat, se kterými se ale při statistických výpočtech setkáváme. Dnes není problém řešit vyrovnaní metodou MNČ přičemž matice plánů (design matrix) mají i několik milionů prvků. V šedesátých letech minulého století to ale problém byl.

Základní myšlenka metody je jednoduchá. Kdybychom statistickým souborem dat o rozsahu m hodnot prokládali polynom $P_n(x)$, měla by matice plánů A rozměry $(m, n + 1)$. Matice $A^T A$ by pak byla čtvercová o rozměrech (m, m) . Rozsah m statistického souboru ale může být skutečně veliký.

Výhodnější proto bude definovat v měřených datech statistického souboru „okno“ o velikosti $(2k + 1)$ hodnot, přičemž musí být splněna nerovnost $2k + 1 > n$, aby došlo k vyrovnaní MNČ. Další modifikace metody MNČ spočívá v tom, že při praktických výpočtech nebudeme úlohu řešit maticově, ale každému bodu definovaného datového „okna“ přisoudíme váhu (konvoluční číslo). A navíc můžeme také využít toho, že v případě prokládání polynomu $P_n(x)$ statistickým souborem měřených dat lze velmi jednoduše vypočítat i jeho p -tá derivace $\partial^{(p)} P_n(x) / \partial x^{(p)}$, kde $i = 0, 1, 2, \dots, n$.

¹ VOŠ a SPŠ Varnsdorf, Mariánská 1100, 40747, Varnsdorf

Tyto váhy jsou předem odvozeny a jsou stanoveny pro volený stupeň polynomu n , šíři datového „okna“ $(2k + 1)$ a řád p derivace prokládaného polynomu podle neznámých koeficientů. Samozřejmě při splnění nerovnosti $p \leq n$. K těmto konvolučním číslům jsou dále stanoveny normalizační faktory v závislosti na nastavení parametrů celé metody. Úloha pak přejde na velmi jednoduché váhování měřených dat podobně jako u klouzavého průměru. Zde se jen jedná o vážený průměr a jednotlivé váhy jsou stanoveny pomocí MNC.

Výpočet pak postupuje tak, že měřenou číselnou hodnotu v centrálním bodě datového „okna“ nahradíme přímo hodnotou vypočtenou. Pak posuneme datové „okno“ o jedno měření dále a postup výpočtu opakujeme. Z hlediska nároků na výpočetní techniku se jedná o velice efektivní záležitost, protože potřebujeme relativně málo paměti a výpočet spočívá jen v cyklech.

Metoda má samozřejmě i nevýhody. Jednak při ní přicházíme o prvních (a posledních) k měřených hodnot a za druhé při nevhodné volbě nastavení metody dochází k zásadnímu zkreslení měřených dat. Tj., filtruje se nejen šum z měření ale i důležité informace.

První zmíněnou nevýhodu můžeme odstranit vhodnou modifikací metody Savitzky-Golay (pro ukázkou bude dále odvozena). Nicméně ta druhá nevýhoda stále zůstává a je třeba metodu skutečně dobře „vyladit“ aby poskytovala vše, co může a zároveň nepůsobila ztrátu informace obsažené v měření.

3. Zobecnění metody Savitzky-Golay

Odvození samotné metody je dlouhé a do tohoto příspěvku by se nevešlo. Proto zde uvedu její obecnější podobu (odvozenou později, než původní metoda), která řeší zároveň problém krajních bodů statického souboru.

Mějme tedy aproximační polynom stupně M v proměnné x :

$$P_M(x) = a_0 + a_1x + a_2x^2 + \dots + a_Mx^M \quad (1)$$

a jemu odpovídající hodnoty $f(-n_L), \dots, f(n_R)$ měřené veličiny. Symboly $-n_L, n_R$ vyjadřují souřadnice měřených hodnot v rámci datového „okna“ vzhledem k jeho centrálnímu bodu. Index L znamená levou část a index P pravou část datového „okna“. Pro centrální bod platí $-n_L = n_R = 0$. Pak g_0 bude hodnota polynomu pro $x = 0$, tedy hodnota a_0 .

Matice plánů A bude mít tvar:

$$A_{ij} = t^j, \quad (2)$$

kde $i = -n_L, \dots, n_R$ a $j = 0, \dots, M$. Systém normálních rovnic pro vektor dx má pak tvar:

$$(A^T A) dx = A^T f \quad (3)$$

A řešení systému normálních rovnic má tvar:

$$dx = (A^T A)^{-1} A^T f \quad (4)$$

Můžeme také zavést formy:

$$\{A^T A\}_{ij} = \sum_{k=-n_L}^{n_R} A_{ki} A_{kj} = \sum_{k=-n_L}^{n_R} k^{i+j} \quad \text{a} \quad \{A^T f\}_j = \sum_{k=-n_L}^{n_R} A_{kj} f_k = \sum_{k=-n_L}^{n_R} k^j f_k \quad (5)$$

A nyní je konvoluční číslo C_n součástí hodnoty a_0 jestliže vektor f nahradíme jednotkovým vektorem e_n , kde $-n_L \leq n < n_R$.

Například pro konvoluční číslo C_{-2} pro bod $n = -2$ (při velikosti intervalu $\{-3; 3\} = 7$ bodů) bude konkrétní tvar vektoru e_n následující:

$$e_n = (0, 1, 0, 0, 0, 0, 0) \quad (6)$$

Vektor C_n konvolučních čísel se spočítá na základě MNČ řešením systému normálních rovnic takto:

$$C_n = (A^T A)^{-1} A^T e_n \quad (7)$$

Jeho složky jsou konvoluční čísla pro všechny řády derivace od 0 do M a pro n -tý bod datového okna. Formálně můžeme tento vzorec pro jednotlivá konvoluční čísla přepsat jako:

$$C_n = \{(A^T A)^{-1} A^T e_n\}_p, \quad (8)$$

kde p je řád derivace pro který konvoluční číslo počítáme. Představuje nám souřadnici hodnoty, která nás zajímá ve vektoru C_n .

Konvoluční číslo můžeme také spočítat takto:

$$C_n = \sum_{m=0}^M \left\{ (A^T A)^{-1} \right\}_{pm} n^m, \quad (9)$$

což je jen jinak zapsán minulý výraz.

Takto odvozená metoda má také dvě velké výhody oproti její klasické formulaci. Konvoluční čísla (váhové faktory) jsou vždy z intervalu (0; 1) a tudíž odpadá potřeba normování normalizačním faktorem. Druhou výhodou je možnost počítat náhradní hodnotu (hodnotu prokládaného polynomu) i pro jiný bod datového „okna“ než je jeho bod centrální. To ovlivňuje volba vektoru e_n .

4. Postup práce při řešení dané problematiky

Ondřej Tůma se tedy pustil ve školním roce 2006 / 2007 do teoretického studia problematiky samotné metody Savitzky-Golay a také do studia přidružených témat z oblastí matematické statistiky, teorie chyb a vyrovnávacího počtu. Samozřejmě včetně metody nejmenších čtverců a jejích vlastností.

To se samozřejmě neobešlo bez mnoha konzultací, při nichž jsme strávili mnoho hodin. Dalším krokem bylo pořízení dat. Studijními materiály byla vysokoškolská skripta „*Teorie chyb a vyrovnávací počet I, II*“ pánů Radoucha a Hampachera, „*Statistické zpracování experimentálních dat*“ pánů Melouna a Militkého a dále „*Přehled statistických metod zpracování dat*“, jehož autorem je Hendl.

Moje vize byla, že student nebude metodu demonstrovat pro obhajobě práce na celostátní přehlídce prací SOČ na umělých datech, ale na reálných měřeních. Zde jsem využil mé známosti s pracovníkem Fyzikálního ústavu Akademie věd ČR, v.v.i., Ing. Mariánem Čerňanským, CSc., který je můj strýc. Ten poskytl mému studentovi reálná data z difrakčního měření na polykrystalických materiálech. Konkrétně se jednalo o rentgenová měření povrchové vrstvy polykrystalického křemíku, která měla tloušťku 15 μ m. Účelem měření bylo určit makroskopická a mikroskopická napětí a velikost krystalických částic (krystalků, krystalitů, krystalových zrn) ve zmíněné vrstvě. Sledovaly se tedy úhlové polohy reflexí na pozadí. Na každém vzorku byly ještě umístěny 4 kontrolní body ze stříbra a monokrystalického křemíku s předem známými polohami reflexí.

Po nastudování problematiky student formuloval metodu Savitzky-Golay ve své zprávě. Popsal velmi precizně teoretické odvození metody a její počáteční nastavení (tedy stupeň aproximačního polynomu, šíři datového „okna“ a řád počítané derivace).

Pro praktické výpočty byl třeba naprogramovat výpočetní aparát. Po konzultaci jsme se dohodli, že tímto výpočetním aparátem bude aplikace psaná pod operačním systémem Win XP a bude psána v programovacím jazyce C++ v interaktivním objektově orientovaném prostředí Borland C++ Builder. Následovala rozvaha datových typů. Zde student správně zvolil pro ukládání vypočtených konvolučních čísel a normalizačních faktorů datový typ „*long double*“ (64 bitů) vzhledem k jejich velikostem (i přes to, že konvoluční čísla i normalizační faktory mají celočíselnou povahu typ „*long double*“ je typem s plovoucí desetinnou čárkou).

Výše zmíněná vlastnost typu „*long double*“ ovšem v praxi znamenala trochu jiný přístup k práci s těmito hodnotami. Tak například není možné testovat rovnost dvou čísel tohoto typu přímo. Je třeba testovat absolutní hodnotu jejich rozdílu a opět nikoliv proti nule, ale proti vámi definované toleranci. Pokud nastavíme toleranci správně, bude zachována ona celočíselná povaha i pro typ „*long double*“.

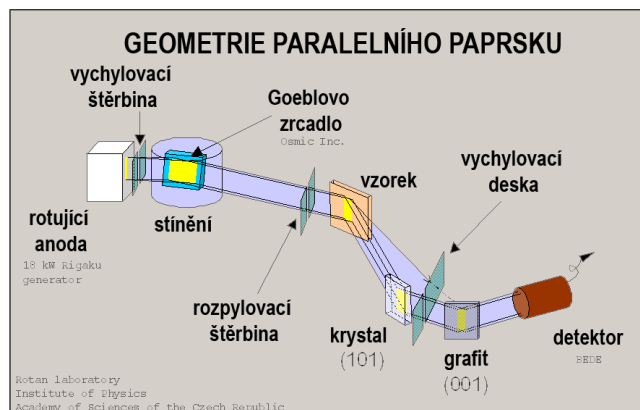
Původní rozvaha počítala s tím, že jeho program bude řešit výpočet samotných konvolučních čísel a normalizačních faktorů, ale také přímo zpracovávat i měřená data. Tato vize se podařila splnit a tak vznikl výpočetní systém na téměř profesionální úrovni.

Součástí práce také byl poster, který prezentoval řešenou problematiku. Tento poster byl také v poster-sekci na konferenci vyvěšen.

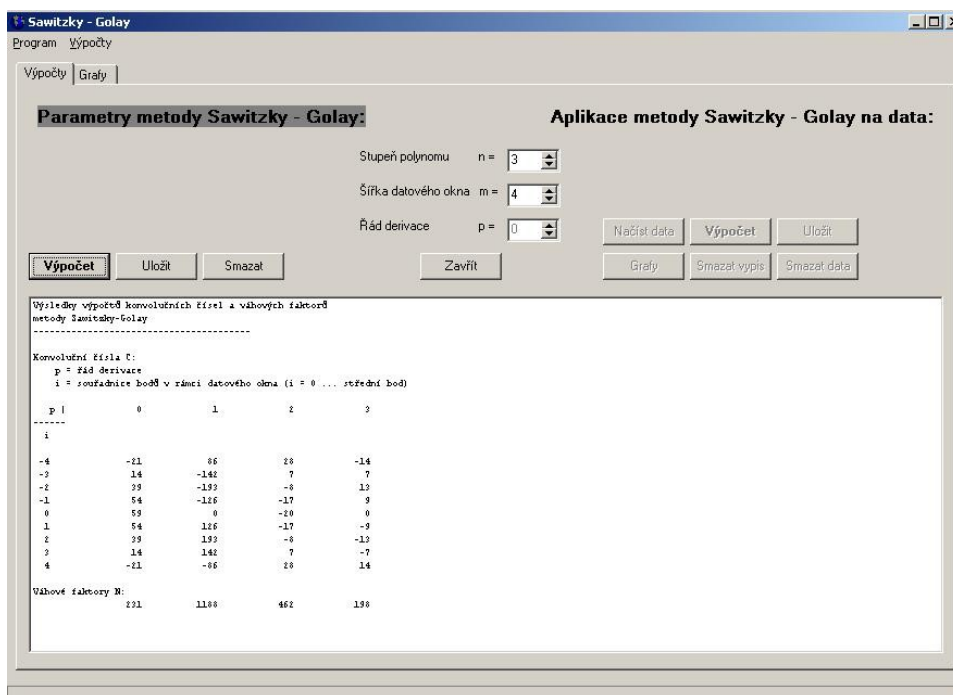
Cíle práce bylo stanovit ideální nastavení parametrů metody Savitzky-Golay tak, aby došlo k filtraci šumu z měření a zároveň aby nedošlo ke ztrátám informací nesených měřeními hodnotami. To se Ondřeji Tůmovi podařilo a jeho výsledky (stupeň polynomu 3., šíře datového „okna“ 9 bodů a řád derivace 3.) plně korespondovaly se zkušenostmi výzkumného týmu Ing. Čerňanského. Je třeba ale předeslat, že ono ideální nastavení této metody jsem znal jen já jako konzultant a tak výsledek, ke kterému student došel zcela sám nebyl mnou nikterak ovlivněn!

Snažení Ondřeje Tůmy bylo korunováno postupem do celostátní přehlídky prací SOČ, kde obsadil 5. místo. Přičemž se na prvních čtyřech místech umístili studenti matematických gymnázií! Považuje tedy jeho výsledek za vynikající.

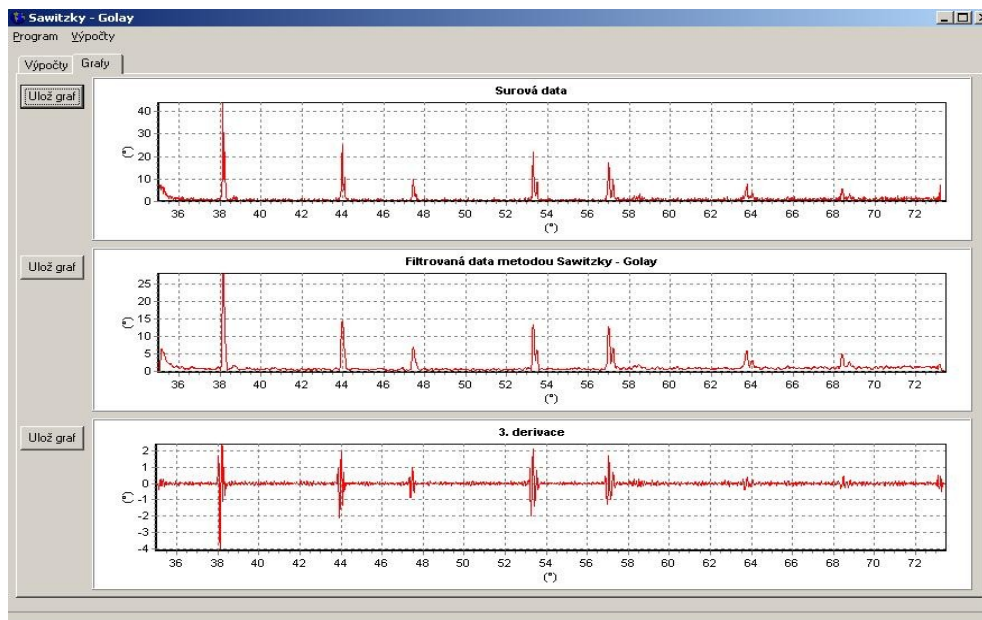
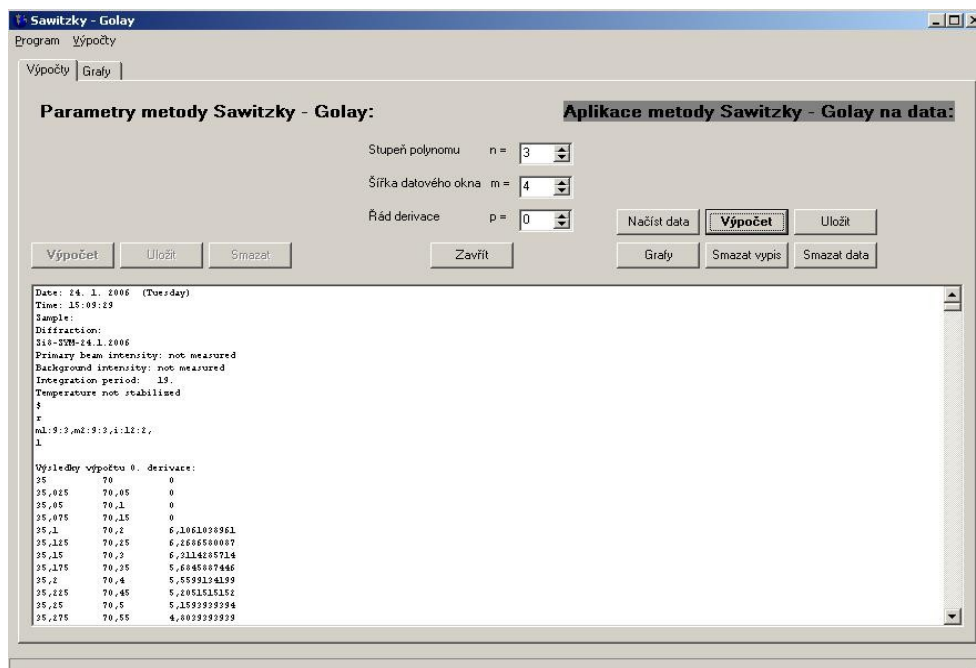
5. Ukázky z práce studenta Ondřeje Tůmy

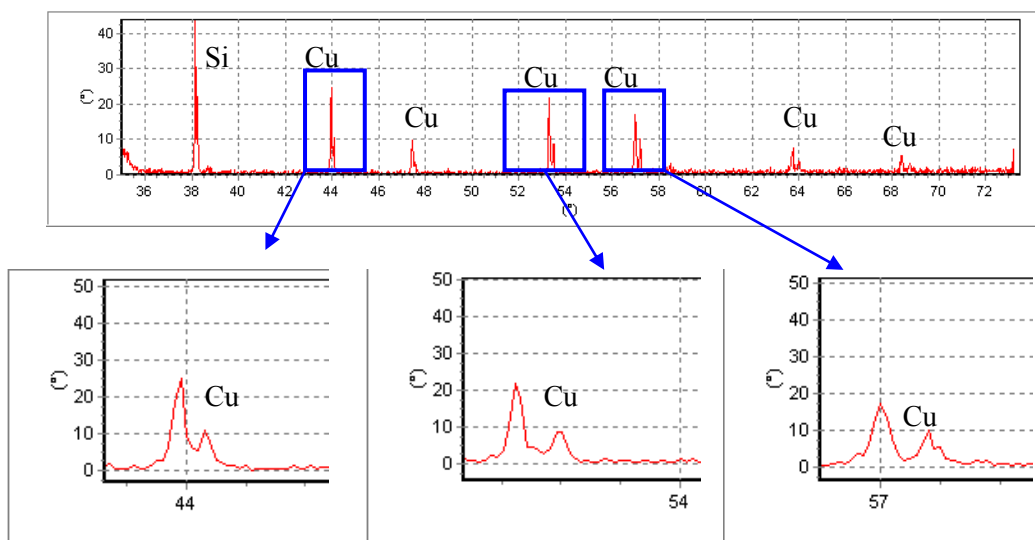


Obrázek 1: Schéma generátoru laserového paprsku a geometrie umístění laboratorního vzorku. Zde student toto schéma převedl do českého jazyka.

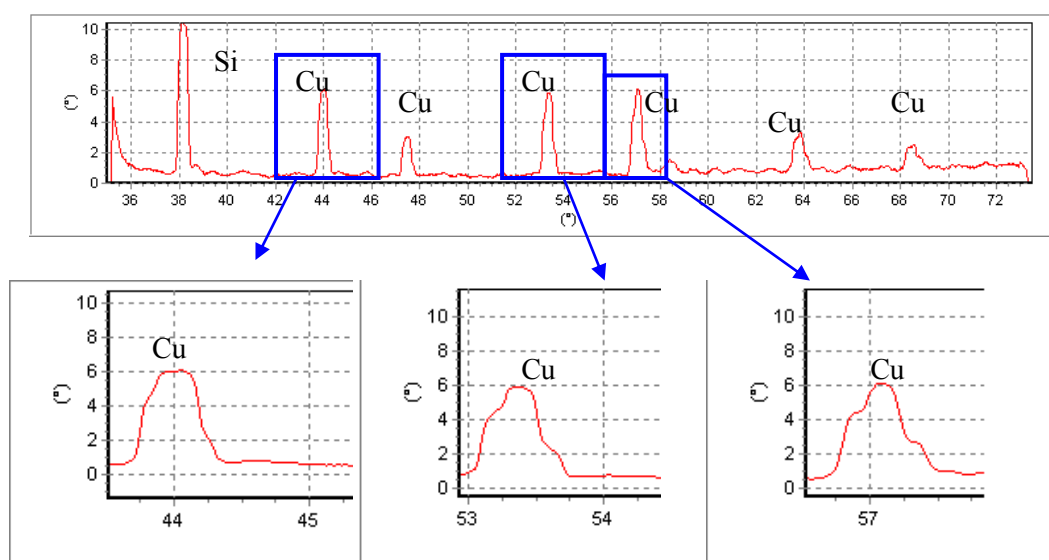


Obrázek 2: Prostředí výpočetního systému, který naprogramoval Ondřej Tůma. Obrázek ukazuje výpočet konvolučních čísel a normalizačních faktorů pro stupeň polynomu 3, šíři datového okna 9 bodů a 0. – 3. řád derivace polynomu

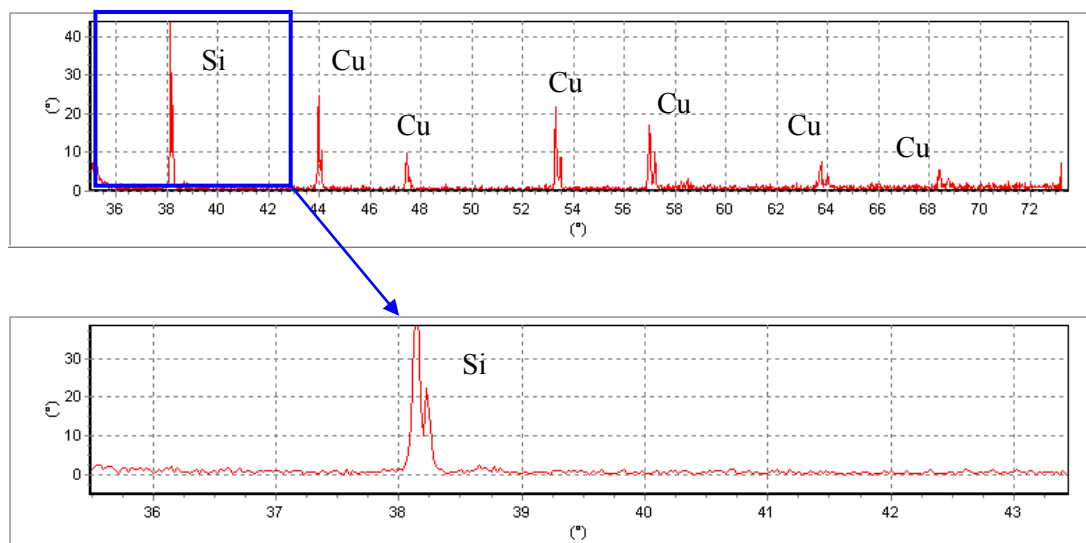




Obrázek 5: Ukázka měřených dat, která ukazují na přítomnost mědi v zkoumaném vzorku polykrystalického křemíku. Pík na 38° je odezvou kontrolního bodu monokrystalického Křemíku.



Obrázek 6: Ukázka filtrovaných dat z obrázku 5 při špatně nastavených parametrech metody. Zde širší datového okna 17 bodů, polynom 1. stupně. Je zde zcela jednoznačně vidět, jak dvojité píky v okolí úhlů 44° , 54° a 57° zmizí a dochází tak k velké ztrátě informací.



Obrázek 7: Ukázka filtrovaných dat z obrázku 5 při správně nastavených parametrech metody. Zde tedy širší datového okna 9 bodů, polynom 3. stupně. Metoda píky zachování a pozadí je čistší než v původních měřených datech.

6. Závěrem

Výsledky práce Ondřeje Tůmy plně korespondují s dobrými zkušenostmi s touto metodou ve Fyzikálním ústavu AV ČR, v.v.i. Sám student skutečně udělal dobrý kus práce a největším přínosem jsou samozřejmě znalosti a dovednosti, které při řešení dané problematiky nasbíral.

Nezbývá než Ondřeji Tůmovi popřát hodně úspěchů ve studiu i v životě. A nám, učitelům, ať už na vysokých či středních školách, popřát více takových schopných studentů se záplem pro věc a snahou proniknout do tajů vědy v kterémkoliv vědním odvětví a odhodlaných věnovat práci a studiu více času ze svého osobního volna. Tedy těch, kteří nejdu tzv. cestou nejmenšího odporu.

7. Literatura

- [1] TŮMA, O. *Metoda Savitzky-Golay*. Práce představená v celostátní přehlídce prací SOČ.
- [2] SAVITZKY, A., GOLAY, M. J. E. *Smoothing and Differentiation of Data by Simplified Least Squares Procedures*, Analytical Chemistry vol. 36, strany 1627-1639, rok 1964
- [3] PRESS W., TEUKLSKY, S. A., *Numerical recipes in Fortran 77: The art of Scientific Computing Vol. 1 of Fortran*, Numerical Recipes, 1996
- [4] GORRY, P. A., *General Least - Squares Smoothing and Differentiation by the Convolution (Savitzky - Golay) Method*, Analytical Chemistry vol. 62, strany 570-573, rok 1990