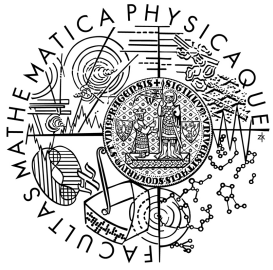
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Jindřich Soukup

Studium počátečních fází růstu kovových vrstev metodami počítačové fyziky

Ústav teoretické fyziky

Vedoucí diplomové práce: prof. RNDr. Rudolf Hrach, DrSc.

Studijní program: Fyzika

Studijní obor: Matematické a počítačové modelování ve fyzice a technice

Praha 2011

Na tomto místě bych chtěl v první řadě poděkovat svému vedoucímu prof. RNDr. Rudolfu Hrachovi, DrSc., jehož poutavé přednášky mě přitáhly k tématu této práce. Chtěl bych mu poděkovat též za vedení této práce, zvláště pak za volnost, kterou mi při zkoumání daného tématu dal. Rovněž mu děkuji za cenné rady a konzultace.
Dále bych zde chtěl poděkovat dvěma lidem, kteří velkou měrou přispěli ke zkvalitnění této práce. Kristýně Kuncové děkuji za odborné i pravopisné korektury práce, svému otci Martinu Soukupovi pak za pravopisné a stylistické úpravy. Zodpovědnost za případné chyby, které v práci přesto zůstaly, jdou na vrub autora.

Prohlašuji, že jsem tuto diplomovou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Podpis autora

Název práce: Studium počátečních fází růstu kovových vrstev metodami počítačové fyziky
Autor: Jindřich Soukup
Katedra (ústav): Ústav teoretické fyziky
Vedoucí diplomové práce: prof. RNDr. Rudolf Hrach, DrSc.
e-mail vedoucího: Rudolf.Hrach@mff.cuni.cz

Abstrakt: Tato práce se zabývá popisem a analýzou obrazových dat, souvisejících s počátečními fázemi růstu tenkých vrstev. Úvodní rešeršní část obsahuje popis tenkých vrstev a způsoby jejich tvorby. Následuje přehled používaných modelů růstu tenkých vrstev. Jádrem práce je pak analýza a modifikace morfologických metod a interpretace jejich výsledků. Důraz je kladen na statistickou povahu metod a jejich optimální implementaci vzhledem k přesnosti výsledku. Práce ukazuje, jak lze modifikovat radiální distribuční funkci a metody založené na Voronoiově dláždění a Delaunayově triangulaci tak, aby lépe postihovaly charakter testovaných dat. Nové metody jsou testovány na experimentálních i modelových datech, sledována je jejich robustnost, citlivost a jejich vzájemná nezávislost. V návaznosti na to je představen a analyzován nový model růstu tenkých vrstev.
Klíčová slova: tenké vrstvy, morfologické metody, Voronoiovo dláždění, Delaunayova triangulace

Title: Computational study of initial stages of metal film growth
Author: Jindřich Soukup
Department: Institute of Physics of the Charles University
Supervisor: prof. RNDr. Rudolf Hrach, DrSc.
Supervisor's e-mail address: Rudolf.Hrach@mff.cuni.cz

Abstract: This work deals with the description and analysis of image data, which related to the initial stages of the thin film growth. The introductory retrieval section includes a description of thin films and methods of their deposition. The following part is an overview of the growth models of thin layers. The heart of my thesis is the analysis and modification of morphological methods and interpretation of their results. The emphasis is placed on the statistical aspect of methods and their optimal implementation due to the accuracy of the results. The work shows how to modify the radial distribution function and methods based on so-called Voronoi and Delaunay triangulation tessellation so that they can better affect the character of test data. New methods are tested both on the experimental and model data. Then we examine their robustness, sensitivity and their mutual independence. At the conclusion it is introduced and analyzed a new model of thin film growth.
Keywords: thin films, morphological methods, Voronoi tessellation, Delaunay triangulation

Obsah

Úvod

Fyzika tenkých vrstev prochází v posledních desetiletích rychlým rozvojem. Její využití v elektronice (mikroprocesory), optice a optoelektronice (antireflexní vrstvy, optické filtry), v dopravě (katalyzátory aut) nebo například v medicíně z ní činí důležitý vědní obor, stimuluje výzkum vlastností takovýchto systémů a vede k vývoji nových metod tvorby tenkých vrstev. Snaha o co největší miniaturizaci pak obrací pozornost k počátečním fázím růstu tenkých vrstev a k jejich vlastnostem.

Geometrie tenkých vrstev a tenkovrstvových systémů se liší podle jejich použití. Tenké vrstvy mohou být připravovány jako dvourozměrné systémy na rovinném podkladu (typická aplikace v mikroelektronice a optoelektronice), jako dvourozměrné systémy na podkladech složitých tvarů (katalyzátory) a nebo jako tzv. kompozitní vrstvy, kdy je jeden materiál, obvykle kov nebo polovodič, rozložen v objemu jiného materiály, obvykle dielektrika (typická aplikace - optické filtry, odporové vrstvy, ale např. i konstrukční materiál raketoplánů). Tato práce se bude zabývat pouze tenkými vrstvami připravovanými na rovinných podkladech.

Teoretický popis tenkých vrstev je z podstaty problému vždy pouze přibližný. Velkou roli zde hrají náhodné vlivy a není jednoduché posoudit, které z nich jsou důležité a které naopak zanedbatelné. Zde nastupují fyzikální modely, které mohou za zjednodušených podmínek vznik tenké vrstvy simulovat. Tyto modely nám mohou poskytnout lepší vhled do problému a ukázat nám procesy, které pomocí měření experimentálních dat nejsme schopni získat. Porovnáním experimentálních dat s výsledky modelů pak můžeme testovat platnost fyzikálních teorií.

Informace o počátečních fázích růstu tenkých vrstev získáváme pomocí snímků z transmisního elektronového mikroskopu. K popisu těchto dat se často používá metod digitálního zpracování obrazu. Zde již existují zavedené a používané metody, stále je ale možné tyto metody zdokonalovat či vynalézat jiné, jak ostatně ukazuje i tato práce.

Vzhledem k tomu, že všechny tři zmiňované obory jsou velice rozsáhlé a není možné je zde v úplnosti popsat, může čtenář v případě hlubšího zájmu sáhnout po další literatuře. V oblasti fyziky tenkých vrstev je možné vřele doporučit knihu [1]. Informace uvedené v kapitole 1 vycházejí převážně z tohoto zdroje. Úvod do počítačové fyziky a zpracování obrazu pak mohou poskytnout například skripta [2] a [3].

Tato práce je rozdělena následovně: První kapitola pojednává o tenkých vrstvách, popisuje základní metody jejich tvorby a průběh tohoto procesu. Obsahuje i fotografie tenkých vrstev z elektronového mikroskopu, které se budou v dalších částech práce analyzovat.

Druhá kapitola obsahuje stručný souhrn existujících modelů tvorby tenké vrstvy. Nastiňuje možnosti modelů a jejich omezení. Obsahuje i výstupy, které jsou v rámci

třetí kapitoly zpracovány.

Aby bylo možné porovnat modelová data s experimentálními, je nutné čtenáře zasvětit do základů teorie pravděpodobnosti a představit některé morfologické metody, které lze aplikovat na obrazová data.

Čtvrtá kapitola obsahuje popis některých statistických metod a ukazuje, jak lze pomocí nich zlepšit morfologické metody z předchozí kapitoly. Zabývá se též výpočtem chyb morfologických metod.

Pátá kapitola propojuje všechny předchozí. Je v ní představen nový model růstu tenkých vrstev. Pomocí morfologických metod jsou pak analyzovány výstupy tohoto modelu.

Šestá kapitola zpětně shrnuje všechny původní výsledky obsažené v této práci. Orientaci v textu může čtenáři ulehčit rejstřík pojmů, umístěný na konci práce.

Kapitola 1

Tenké vrstvy

Podle definice je tenká vrstva pevnou látkou, jejíž alespoň jeden rozměr je velmi malý natolik, že se mění objemové vlastnosti látky. Tato definice ale není zcela obecná, protože se může vztahovat k různým fyzikálním vlastnostem, kterými tenkou vrstvu charakterizujeme - jejím vlastnostem mechanickým, optickým, elektrickým, apod. Zatímco podle jedné vlastnosti se materiál ještě chová jako objemový, podle jiné již jako tenká vrstva. V případě, že máme strukturu se dvěma rozměry nekonečnými (resp. makroskopickými) a jedním konečným, typická hranice tloušťky tenké vrstvy leží v rozmezí několika desítek nanometrů až mikrometru. V případě nespojité nebo polospojité vrstvy stejně jako v případě kompozitního materiálu se vždy jedná o tenkou vrstvu.

1.1 Experimentální metody

K vytváření tenkých vrstev se používá velké množství metod. Každá z nich má své výhody i svá omezení. Tato práce bude využívat experimentální data vytvořená pomocí vakuového napařování. Pro přehlednost se však pokusíme shrnout i ostatní metody.

Metody je možné rozdělit do dvou skupin - na „chemické a elektrochemické“ a na „fyzikální“. Dá se říci, že větší uplatnění mají fyzikální metody. U mnohých metod neexistují pro jejich názvy české ekvivalenty. V takových případech budeme užívat anglických názvů.

1.1.1 Chemické metody

Mezi nejdůležitější chemické a elektrochemické metody patří tyto:

- **Cathode electrolytic deposition** - u této metody je látka, ze které má být vytvořena tenká vrstva, rozpuštěná v roztoku nebo roztavená, v obou případech ve formě iontů. Pokud vložíme do roztoku (či taveniny) elektrody, kladné ionty kovů budou přitahovány ke katodě, kde se po přijetí elektronu usadí. Hmotnost nanesené látky bude úměrná elektrickému náboji, konstantou úměrnosti je elektrochemický ekvivalent daného materiálu. Vlastnosti výsledné tenké vrstvy závisejí na složení elektrolytu. Při této metodě je možné jako substrát (podklad)

použít jen vodivé materiály. Tenké vrstvy mohou být znečištěny dalšími složkami roztoku.

- **Electroless deposition** - tato metoda je založená na stejném principu jako předchozí. V tomto případě však za tvorbu tenké vrstvy nemůže externí elektrický zdroj, ale elektrochemické procesy probíhající na elektrodách. Rychlost tvorby tenké vrstvy závisí na teplotě roztoku, často je potřeba použít nějaký katalyzátor. Tato metoda se používá například při poniklování.

- **Anode oxidation** - tato metoda je používána hlavně k tvorbě tenkých vrstev oxidů některých kovů (Al, Ta, Nb, Ti, Zr, ...). Toto vrstvy se tvoří oxidací anody, ke které jsou díky silnému elektrickému poli přitahovány ionty kyslíku. Ty pak mohou difundovat skrz již vytvořenou vrstvu oxidu. Rychlost růstu závisí exponenciálně na intenzitě elektrického pole. Tloušťka vrstvy je díky tomu velice homogenní, nelze však vytvářet příliš silné vrstvy, v určitém okamžiku dochází k elektrickému průrazu materiálu.

- **Chemical vapor deposition** - je často používána při výrobě polovodičových součástek. Je možné dosáhnout vysoké čistoty a krystalického charakteru vrstvy. Často se takto připravují homeoepitaxní vrstvy (podložka je ze stejného materiálu jako vznikající tenká vrstva). V metodě se využívá určitých typů chemických reakcí jako je pyrolýza (rozpad za vysokých teplot) nebo fotolýza (rozklad způsobený ultrafialovým či infračerveným zářením) některých plynných sloučenin. Příkladem jsou hydridy GeH_4 či SiH_4, ze kterých po rozkladu získáme čisté germanium či křemík, které pak vytvářejí na podložce tenké vrstvy.

Nevýhoda chemických metod je, že jsou použitelné vždy jen na malou skupinu materiálů. Výhodou je velká čistota získaných vrstev. Chemické metody jsou často s úspěchem využívány v polovodičové technice.

1.1.2 Fyzikální metody

Hlavní dvě fyzikální metody jsou **vakuové napařování** (vacuum evaporation) a **katodové naprašování** (cathode sputtering). U obou metod existuje mnoho modifikací (laser beam evaporation, electron bombardment, resp. magnetron sputtering). Obě metody se vyznačují značnou univerzalitou, co se týká nanášených materiálů. Vždy je potřeba speciální aparatura a přesně definované pracovní prostředí - vakuum v první případě, prostředí vyplněné inertními plyny v druhém.

Proces tvorby tenké vrstvy pomocí **vakuového napařování** probíhá následujícím způsobem. Nejprve se dodáním energie vypaří či vysublimuje materiál, který chceme napařit. Odpařené částice následně putují k substrátu. Po dopadu se částice mohou či nemusí zachytit. Může docházet k driftování částic po povrchu, opětnému odpaření nebo naopak úplnému zachycení.

Jednotlivé varianty této metody se liší zejména tím, jak je dodáváno teplo materiálu, který chceme vypařit (vysublimovat). V praxi se nejčastěji používá ohmické teplo či odpařování elektronovým dělem nebo laserovým paprskem. Důležitým parametrem je tlak plynů uvnitř aparatury. Molekuly plynů mohou rozptýlit proud částic

putujících k substrátu a naopak se mohou začleňovat samy do vznikající tenké vrstvy, což je v naprosté většině případů nežádoucí. Proto se uvnitř aparatury udržuje vysoké vakuum, řádově 10^{-6} Pa. O jevech probíhajících po dopadu napařovaného materiálu na substrát bude pojednáno v následující části.

Při **katodovém naprašování** je z naprašovaného materiálu vytvořena katoda systému, v němž dochází vlivem silného elektrického pole k doutnavému výboji. Substrát, na kterém vytváříme tenkou vrstvu, je umístěn mezi katodou a anodou systému. Pracovním prostředím bývá některý z inertních plynů, například argon či xenon. Kladné ionty těchto plynů vzniklé výbojem jsou pak přitahovány ke katodě. Při jejich dopadu se mohou uvolnit částice (neutrální atomy, ionty, občas i větší shluky částic) materiálu, ze kterého je katoda vytvořena. Tyto částice kondenzují na okolním povrchu a tedy i na substrátu.

Existuje mnoho způsobů, jak výše popsanou základní konfiguraci vylepšit. Cílem je zejména snížení tlaku vzácných plynů, které umožní zvýšit čistotu vzniklé vrstvy.

1.2 Mechanismus vzniku tenkých vrstev

V závislosti na tom, jak velké síly působí mezi atomy napařovaného materiálu a substrátem a mezi atomy napařovaného materiálu navzájem, můžeme rozlišovat tři módy tvorby tenkých vrstev:

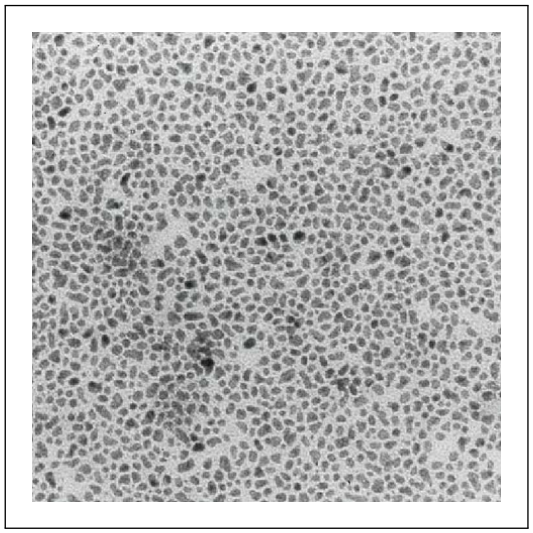
- Van der Merweův mechanismus - tvoří se postupně jednolité vrstvy.
- Volmerův-Weberův mechanismus - tvorba nukleačních center, tvorba, růst a spojování trojrozměrných ostrůvků.
- Stranského-Krastanovův mechanismus - adsorbce jedné nebo více monovrstev, na nichž se tvoří trojrozměrné ostrůvky.

Blíže se budeme zabývat Volmerovým-Weberovým mechanismem, který je typický pro růst kovových vrstev na dielektrických podložkách.

V tomto případě lze rozeznat celkem čtyři fáze růstu tenké vrstvy:

1. **Tvorba nukleačních jader** - první částice jsou adsorbovány na povrch buď vlivem chemické adsorbce nebo při setkání s další částicí.
2. **Růst jader** - atomy migrující po povrchu se nabalují na nukleační jádra. Ostrůvky rostou, často dochází ke krystalizaci.
3. **Slévání ostrůvků** - vytváří se polospojitá a následně spojitá vrstva.
4. **Tloušťkový růst**

Jednotlivé fáze nejsou časově oddělené, k tvorbě nových nukleačních jader (tzv. sekundární nukleaci) může docházet i během druhé a třetí fáze.

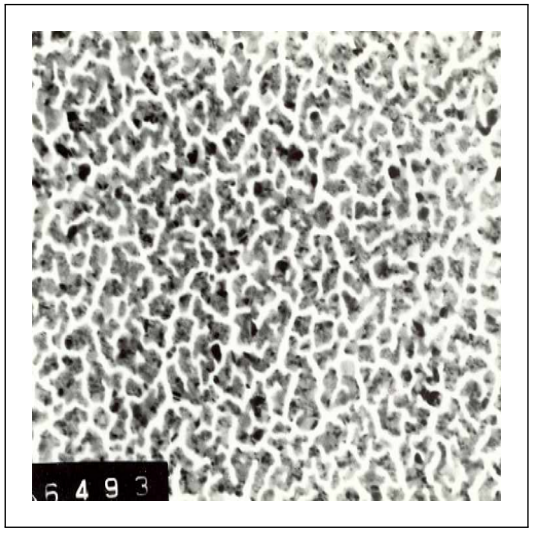


Obrázek 1.1: Fotografie zachycující počáteční fázi tenkých vrstev rostoucích Volmerovým-Weberovým mechanismem [4]

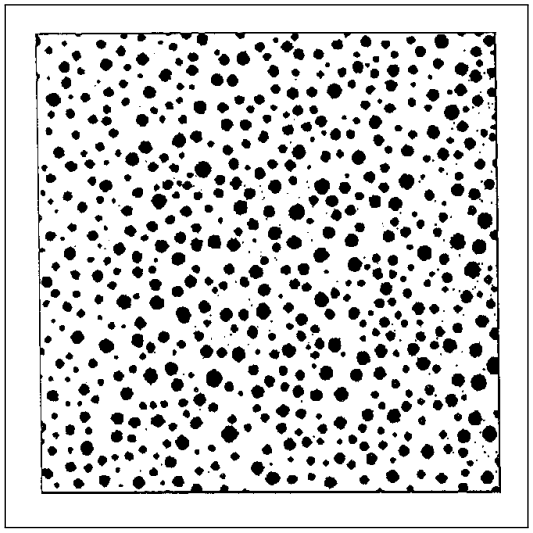
1.3 Vymezení

V této práci se zaměříme pouze na počáteční fáze růstu tenkých vrstev Volmerovým-Weberovým mechanismem. Pro tyto systémy lze získat pomocí transmisního elektronového mikroskopu kvalitní fotografie, viz obrázky 1.1, 1.2 a 1.3.

První obrázek nám dává představu o měřítkách, ve kterých se při tvorbě tenkých vrstev pohybujeme. Na obrázku je výřez fotografie o velikosti cca 2×2 cm při zvětšení elektronového mikroskopu $100\,000\times$. Hrana zachycené části vzorku je tedy dlouhá přibližně 200 nm. Druhý obrázek ukazuje, jak může vypadat polospojitá vrstva naneseného materiálu. Třetí obrázek ukazuje, jak vypadá snímek tenké vrstvy po tzv. binarizaci. Binarizace je proces, při němž dochází k rozdělení fotografie na dvě části - pozadí a objekty - přičemž objekty jsou označeny černě a pozadí bíle. V této práci budeme pracovat pouze s binarizovanými fotografiemi.



Obrázek 1.2: Ukázka polospojité struktury [4]



Obrázek 1.3: Fotografie tenké vrstvy stříbra připravené na dielektrické podložce – binarizovaný snímek [5]

Kapitola 2

Modely růstu tenkých vrstev

Fyzikální model je postup či myšlenková konstrukce, která má za cíl zjednodušeně popsat daný fyzikální jev. V případě modelů růstu tenkých vrstev můžeme způsoby zjednodušení rozdělit do několika skupin.

První kategorií je prostorové omezení. Musíme si určit, jestli do modelu zahrneme celou aparaturu nebo jen malou část podložky, na které vzniká tenká vrstva. Musíme se také rozhodnout, jestli budeme pracovat v trojrozměrném prostoru, nebo se omezíme na dvojrozměrnou projekci.

Druhou kategorií je časové omezení. Některé modely časové měřítko úplně opomíjejí a snaží se modelovat stav v jednom konkrétním okamžiku. Jiné modely s časem pracují a snaží se modelovat kratší či delší časový úsek procesu.

Třetí kategorií zjednodušení jsou aproximace objektů vyskytujících se v experimentu. Jedním extrémem je popisovat objekty v experimentu pomocí vlnové funkce a k propagaci využívat Schrödingerovu rovnici. Opačným extrémem je odpoutání objektů od jejich fyzikální podstaty. Pak můžeme například ostrůvky vznikající u Volmerova-Weberova mechanismu reprezentovat množinami pixelů na obrázku.

Cílem modelování je získat informace o experimentu, aniž bychom ho museli provádět. Abychom mohli považovat tyto informace za věrohodné, musíme nějakým způsobem ověřit, jestli a v jakých ohledech model odpovídá experimentu. V případě tenkých vrstev můžeme k porovnání použít fotografie tenkých vrstev z elektronových transmisních mikroskopů, případně dalších zobrazovacích technik (STM, AFM, atd.).

Pokud bychom porovnali výsledky modelu s realitou, zjistili bychom, že výsledky modelu se od reality více či méně odlišují. My bohužel realitu přesně neznáme. Informaci o ní nám zprostředkovávají fotografie z elektronového mikroskopu, které mají často nízké rozlišení a na nichž v některých místech ani nejsme schopni rozpoznat, co jsou napařené atomy a co podložka. Při porovnávání fotografií s výsledky modelů tedy porovnáváme nepřesné s nepřesným. Pokud bychom dokázali vytvořit model lišící se od fotografií z elektronového mikroskopu srovnatelně tomu, jak se liší fotografie od skutečnosti, mohli bychom z modelu teoreticky získat kvalitnější informace než z experimentu.

Modely, které dostatečně dobře popisují experiment, můžeme využít i ke zkoumání fyzikálních jevů samotných. V experimentech jsme omezeni dostupnou aparaturou, materiály a jejich vlastnostmi. V simulacích můžeme libovolně měnit vlastnosti materiálů

a tím zkoumat jejich vliv na výslednou strukturu. Modely nám tedy mohou pomoci lépe pochopit fyzikální zákonitosti daného fyzikálního jevu. V praxi se k tomuto účelu často používají i modely značně nepřesné. Závěry vytvořené na základě simulací je tedy nutné vždy brát s rezervou a porovnávat s výsledky experimentálními.

Ve zbytku této kapitoly popíšeme některé často využívané modely tenkých vrstev. Vysvětlíme si jejich princip, nastíníme jejich omezení a přednosti, a uvidíme, jaké informace můžeme z těchto modelů získat. Tento souhrn má za cíl dát čtenáři prvotní představu o tom, jaké modely se v této oblasti používají. Druhým cílem této kapitoly je popsat modelová data, se kterými budeme v dalších kapitolách pracovat. Metodami, kterými lze porovnávat modelová a experimentální data, se zabývá kapitola 3.

2.1 Hard-disk model

Jeden z nejjednodušších modelů se nazývá hard-disk model.[1] Model je čistě stochastický (není deterministický, využívá (pseudo)náhodných čísel), nemá vazbu na fyzikální podklad a pracuje s dvojrozměrnou projekcí.[2] Růst tenkých vrstev napodobuje značně nepřesně, bývá však často používán k testování metod pracujících s obrazovými daty [8, 9] či k porovnání se složitějšími modely [5, 10].

Model bere ostrůvky vznikající Volmerovým-Weberovým mechanismem jako celky a nezabývá se jejich vnitřní strukturou. Ostrůvky aproximuje kruhy s předem určeným poloměrem R [3]. Při simulaci se jednotlivé kruhy náhodně umísťují do pracovní oblasti $\Omega \subset \mathbb{R}^2$. Výstupem modelu je množina N objektů umístěných v oblasti Ω (objekt je umístěný v oblasti Ω, pokud tam leží jeho těžiště). Jediným omezením při umísťování objektů je, aby neležely příliš blízko u sebe. Parametr, který určuje nejbližší možnou vzdálenost dvou objektů od sebe, se nazývá **difúzní parametr** D. Jako vzdálenost objektů zde bereme vzdálenost mezi jejich okraji.

Posledním parametrem modelu je tzv. **poměr zaplnění** P. Pohybuje se mezi nulou a jedničkou a značí, kolik objektů bylo na plochu vloženo v poměru k maximálnímu možnému počtu objektů N_{max}, které tam bylo možné v daném případě vložit.

2.1.1 Praktická realizace modelu hard-disk

Maximální počet objektů, které lze na pracovní plochu umístit, není vždy konstantní a pro jednotlivé realizace modelu se může lišit až o několik procent. Pokud jako parametr modelu bereme poměr zaplnění, musíme při simulaci zjistit i veličinu N_{max} příslušnou dané konkrétní realizaci. K tomu používáme následující algoritmus:

1. **Počáteční fáze.** V první fázi generujeme souřadnice, na které se pokoušíme umístit objekt, náhodně. První fázi ukončíme v okamžiku, kdy už je pravděpodobnost umístění velmi nízká.

[1] Zde využíváme názvosloví používané prof. Hrachem a jeho spolupracovníky [5, 6]. Tento termín je inspirován variantou molekulárně-dynamické simulace s obdélníkovým potenciálem. Při vhodném nastavení parametrů lze tento model také popsat matematicky jako hard core point process [7].

[2] Existuje i trojrozměrná varianta, která se používá při modelování kompozitních materiálů.

[3] V některých případech jsou ostrůvky aproximovány i jinými geometrickými tvary, například elipsami [8].

2. **Primární dělení.** Plochu si rozdělíme na $p_1 \times q_1$ obdélníků. U každého obdélníku otestujeme, zda je možné, aby tam mohl ležet střed nového kruhu, a to následujícím způsobem: testujeme, zda existuje objekt, jehož těžiště je blíže středu obdélníku, než je $D + 2R - q$. Hodnota q zde značí polovinu úhlopříčky obdélníku. Pokud takový objekt existuje, není možné vložit na plochu objekt, který by měl těžiště v tomto obdélníku.

   Toto tvrzení nemusí být na první pohled zřejmé, proto se k němu později vrátíme a objasníme si ho podrobněji.

3. **Druhá fáze.** V tomto kroku se pokoušíme umístit objekty jen do obdélníků, které jsme v předchozím kroku nezamítli. Druhou fázi ukončíme, když pravděpodobnost umístění klesne k nule, resp. když dlouhou dobu neumístíme žádný objekt.

4. **Sekundární dělení.** Každý obdélník z druhé fáze rozdělíme na $p_2 \times q_2$ menších obdélníků a každý z nich otestujeme stejně jako v kroku dva. S obdélníky, které jsme nevyloučili, se vracíme do bodu tři.

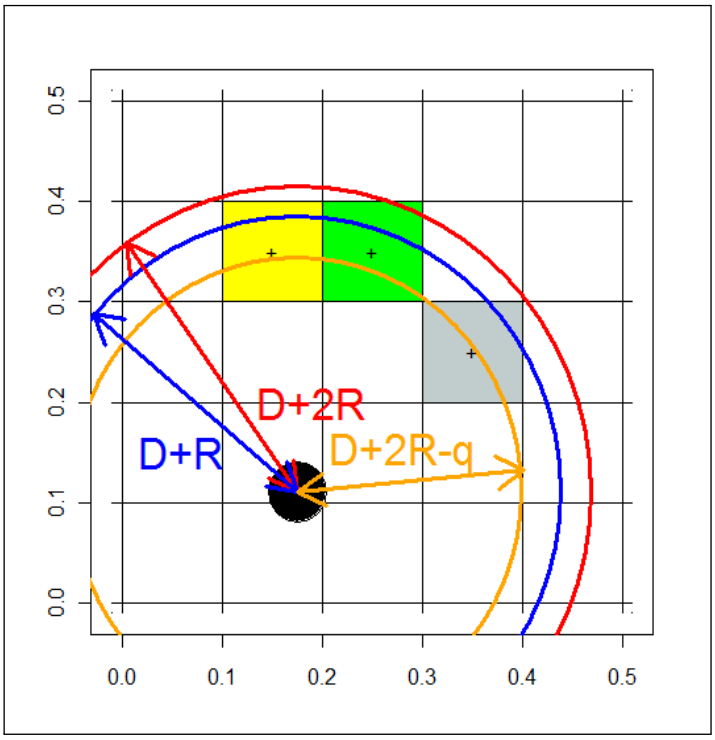
Pokud jsme v druhém či čtvrtém kroku algoritmu vyloučili všechny obdélníky, ukázalo se nám, že už žádný další objekt na plochu vložit nemůžeme a zjistili jsme též N_{max} příslušné této realizaci. Abychom získali realizaci modelu s příslušným poměrem zaplnění, stačí pak odstranit příslušný počet naposledy přidaných objektů.

Je teoreticky možné, že se algoritmus opakovaným dělením dostane až k obdélníkům s hranou srovnatelnou se strojovou přesností a přesto zbydou nevyloučené obdélníky. Tato možnost je ale velice málo pravděpodobná. Při testování algoritmu se při nastavení dostatečně dlouhé první a třetí fáze algoritmus zastavil téměř vždy po prvním sekundárním dělení (byly vyloučeny všechny obdélníky prvního sekundárního dělení).

Nyní se vrátíme k druhému kroku algoritmu a ilustrujeme si ho na situaci naznačené na obrázku 2.1. Na něm je naznačeno rozdělení plochy na obdélníky. Na ploše je umístěn jeden objekt a okolo něho jsou naznačeny vzdálenosti $D + R$, $D + 2R$ a $D + 2R - q$ od jeho středu.

Podívejme se, jak algoritmus v druhém bodě vyhodnotí vybarvené obdélníky. Zeleně vybarvený obdélník algoritmus nezavrhne, jelikož jeho těžiště je až za hranicí oranžové kružnice. Vidíme, že další objekt do tohoto obdélníku opravdu můžeme umístit (kruh ohraničený červenou kružnicí ho nezakrývá celý). Šedý obdélník algoritmus zavrhne (jeho těžiště leží uvnitř oranžového kruhu). Zajímavá situace nastává u žlutého obdélníku. Červená kružnice nám ukazuje, že do tohoto obdélníku už žádný další objekt umístit nemůžeme, přesto ho algoritmus v tomto kroku nevyloučí. Stane se tak až později, když se algoritmus vrátí zpět do druhého kroku se zjemněnou sítí. Kritérium v druhém kroku tedy vždy propustí ty obdélníky, do kterých objekty umístit ještě lze, nezamítne však nutně všechny obdélníky, kam objekt umístit nelze.

Nyní se podívejme na konkrétní výstupy modelu. Na obrázku 2.2 vidíme výsledky modelu pro oblast $\Omega = [0; 1] \times [0; 1]$, $R = 0{,}005$, $D = 0{,}01$ a $P = \{\frac{1}{6}, \frac{2}{6}, \frac{3}{6}, \frac{4}{6}, \frac{5}{6}, 1\}$. Vidíme, že při vyšším poměru zaplnění se zdá výsledná struktura pravidelnější a objekty jsou rovnoměrněji rozmístěné. Měřením těchto vlastností se budeme zabývat v příští



Obrázek 2.1: Ilustrace algoritmu používaného v modelu hard-disk

kapitole. V algoritmu jsme použili $p_1 = q_1 = 1000$, $p_2 = q_2 = 10$. Výsledný počet objektů byl $N_{max} = 1802$.

Tento model budeme v následující kapitole používat k testování jednotlivých morfologických metod. Z modelu hard-disk vychází i model, který představíme v kapitole 5.

2.2 Molekulárně-dynamické modely

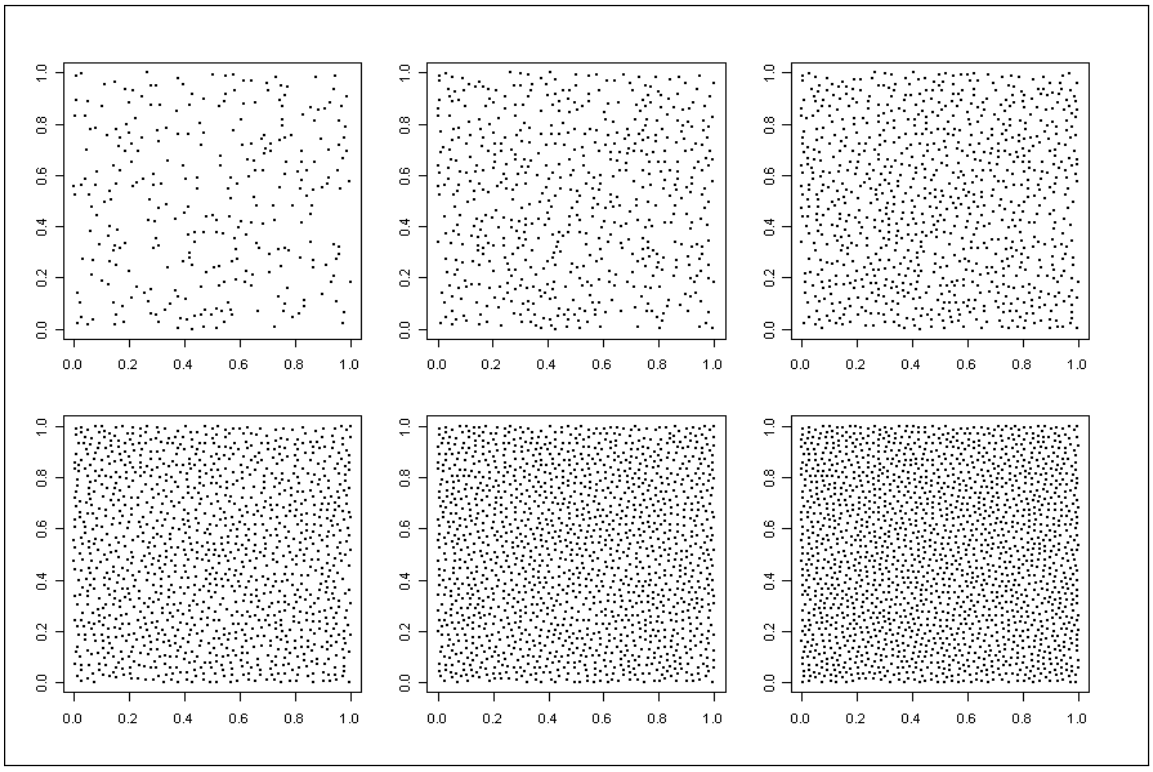
Molekulárně dynamické modely (dále zkráceně MD modely) se v současné době s úspěchem používají v různých odvětvích fyziky, chemie i biologie. Pracují na atomární úrovni v trojrozměrném prostoru a využívají klasické pohybové rovnice, které řeší numericky. Působení sil mezi jednotlivými atomy je zde aproximováno pomocí empirických potenciálů.

Výsledky MD modelů bývají často velice přesné. Problémem je však velká výpočetní náročnost a extrémně malý časový krok (řádově 10^{-15} s)[4]. V závislosti na použitém algoritmu může být výpočetní složitost nejlépe lineární s počtem atomů a lineární s časem. Při současném stavu výpočetní techniky je možné provádět simulace řádově 10^6 atomů po dobu řádově 10^{-6} s, a to jen na velkých počítačových clusterech.[5]

Proces tvorby tenkých vrstev probíhá v řádech sekund a podložka, na kterou na-

[4]Časový krok musí být kratší, než je perioda nejrychlejšího pohybu, který chceme zachytit. V tomto případě jde o vibrace atomů.

[5]Limitní hodnoty vycházejí z informací v článku [11] z roku 2009.



Obrázek 2.2: Realizace hard-disk modelu pro různé poměry zaplnění

pařujeme atomy má řádově víc než 10^6 atomů. Při simulacích tedy musíme rapidně omezit studovanou oblast a/nebo se omezit na velice krátké časové intervaly. Proto lze tímto způsobem simulovat jen určité fragmenty procesu. V článku [12] byly například tímto způsobem zkoumány tvary ostrůvků o různém množství atomů, proces slévání blízkých ostrůvků a také migrace jednotlivých atomů po podložce. Tyto výsledky pak byly použity v metodě Monte Carlo.

2.3 Modely založené na metodě Monte Carlo

Metoda Monte Carlo (dále budu značit jako MC) je názvem pro skupinu stochastických algoritmů použitelných pro velkou škálu problémů. Typickým postupem u těchto modelů je nahrazení deterministických pravidel, jejichž vliv lze postihnout nějakou veličinou, náhodným procesem se stejným rozdělením. Namísto aplikování původního, často výpočetně náročného pravidla, MC modely náhodně generují výsledek tak, aby rozdělení měřitelných veličin zůstalo co nejpodobnější. Typickým příkladem takového modelu je aproximace Brownova pohybu náhodnou procházkou. Místo abychom uvažovali všechny možné síly působící na pylové zrno, změříme si pouze difúzní koeficient. Následně opakovaně generujeme (pseudo)náhodná čísla s vhodným rozdělením, které budou určovat, jakým směrem a jak daleko se pylové zrno v každém časovém kroku posune. Při použití vhodného rozdělení při generování náhodných čísel můžeme Brownův pohyb modelem velmi dobře aproximovat.

Modely typu MC, simulující růst tenkých vrstev, můžeme najít například v práci [13] či v již zmíněném článku [12]. Základními objekty v simulaci jsou atomy a jejich pohyb je simulován pomocí statistik získaných z molekulárně-dynamických simulací. Výsledky takovéhoto modelu mohou kvalitativně vystihovat chování reálných fyzikálních systémů [13]. Díky tomu můžeme pomocí MC modelů zkoumat vliv jednotlivých parametrů experimentu/simulace, například vliv rychlosti napařování, jak ukazuje článek [12].

Nyní stručně popíšeme model z článku [12], jehož výsledky budeme, stejně jako modelová data vygenerovaná hard-disk modelem, využívat v kapitole 5. Jak již bylo řečeno, model přebírá výsledky MD modelu, který simuluje chování jen malé části experimentu. Převzaté informace zahrnují velikosti ostrůvků v závislosti na počtu atomů, popis slévání blízkých ostrůvků, popis migrace jednotlivých atomů po podložce a tvorby kondenzačních jader - zárodků budoucích ostrůvků. Samotný MC model v každém kroku vyhodnocuje, který z objektů na ploše se posune, přidává nové dopadnuvší atomy rychlostí úměrnou napařovací rychlosti a řeší spojování dostatečně blízkých objektů.

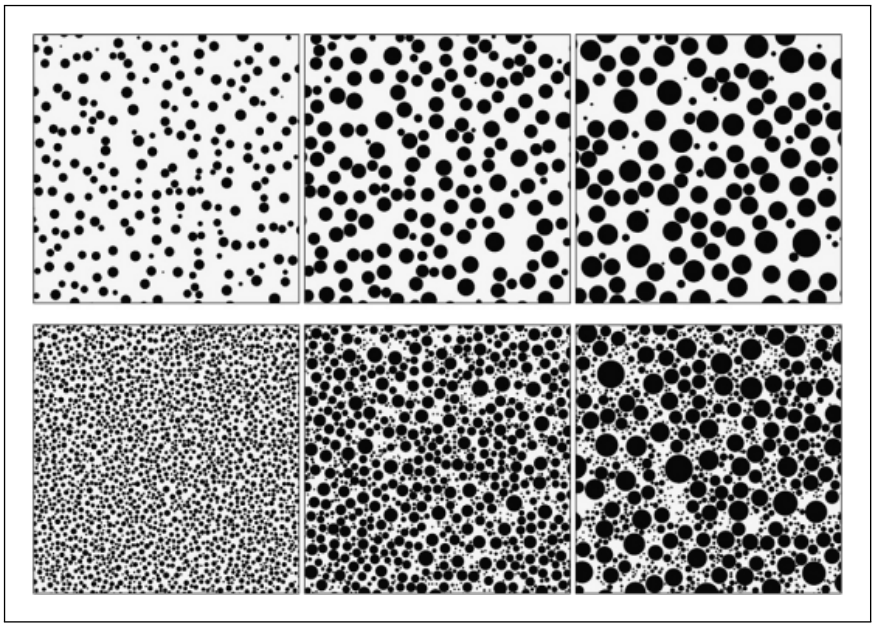
Pracovní oblast je v modelu rozdělena pomocí čtvercové mřížky na 1000×1000 čtverců, těžiště objektů mohou ležet jen v uzlových bodech této mřížky. Ostrůvky i jednotlivé atomy jsou aproximovány kruhovými objekty s příslušným poloměrem převzatým z MD modelu, jsou tedy plně popsány polohou těžiště a poloměrem. V celém systému se mohou pohybovat pouze jednotlivé atomy. Pokud se dva takové atomy potkají, vytvoří dvojici (nukleační jádro) a nadále už zůstanou na místě. Pokud narazí pohybující se atom na skupinu atomů, je jí absorbován a u výsledného objektu se zvětší poloměr kruhu, který jej reprezentuje. Pokud se u nějakého objektu zvětší poloměr natolik, že se dotýká či překrývá s jiným objektem, jsou tyto objekty sloučeny v jeden větší.

Výstupem modelu jsou pozice objektů na ploše a jejich poloměr v různých časových okamžicích. Na obrázku 2.3 vidíme stav systému pro tři časové okamžiky a pro dvě napařovací rychlosti. Napařovací rychlost je definována tím, kolik monovrstev (monolayer, zkratka ML) by se z atomů, které dopadnou na povrch za jednu sekundu, vytvořilo. Odpovídající jednotkou je monovrstva za sekundu (zkratka ML/s). Časové okamžiky, ve kterých jsou systémy zachyceny, jsou takové momenty, kdy dopadlo na povrch tolik atomů, aby vytvořily právě 1, 3, resp. 5 monovrstev.

Z obrázku vidíme, že charakter výsledků modelu je silně závislý na napařovací rychlosti. Jedno z vysvětlení rozdílů může spočívat v tom, že při nízké napařovací rychlosti migrující atomy častěji narazí na větší objekt než na jiný migrující atom. Vzniká tak relativně málo ostrůvků, které ale o to rychleji rostou. Pokud je naopak napařovací rychlost vysoká, migrující atom mnohem častěji narazí na jiný migrující atom a vytvoří další nukleační jádro.

Popsaný model má řadu volitelných parametrů, které mohou pomoci přesněji simulovat konkrétní experimentální podmínky - lze uvažovat i omezenou migraci malých několikaatomových objektů, rozpad objektů a jejich desorpci, rozdílné koeficienty kondenzace nanášených atomů na čisté podložce a površích již vytvořených objektů, atd. Data používaná v této práci však odpovídají zjednodušenému modelu z článku [12].

Detailní informace o tomto modelu vycházejí jak z článku [12], tak i ze soukromé



Obrázek 2.3: Výsledky MC modelu pro napařovací rychlosti 10^2 ML/s (nahoře) a 10^4 ML/s (dole) pro tloušťku vrstev 1 ML (vlevo), 3 ML (uprostřed) a 5 ML (vpravo), výřez [12]

komunikace s prof. Hrachem, jenž je hlavním autorem článku. Profesor Hrach též poskytl data, která budeme používat v kapitole 5.

Kapitola 3

Morfologické metody

Pokud chceme určovat přesnost nějakého modelu, musíme si stanovit kritéria, podle kterých budeme přesnost posuzovat. Od modelů očekáváme, že jejich výsledek bude podobný experimentálně získaným datům - snímkům struktury z transmisního elektronového mikroskopu. Vzhledem k tomu, že snímky různých částí struktury nejsou totožné, není cílem modelu dosáhnout přesně stejné struktury, ale spíše napodobit charakter snímků. K popisu vlastností obrázků budeme používat buď nějaké funkční závislosti, kterým budeme říkat **charakteristiky**, nebo číselné hodnoty, tzv. **příznaky**. Metodu, pomocí které získáme daný příznak či charakteristiku, budeme nazývat **morfologickou metodou**.

V této kapitole představíme několik morfologických metod a příznaků z nich odvozených. Dále budeme testovat tyto metody na modelových datech představených v předchozí kapitole. Ukážeme, jak lze tyto metody modifikovat pro zpracování experimentálních dat a zhodnotíme jejich vhodnost pro fotografie různých stádií růstu tenkých vrstev.

Kvalitní zhodnocení vlastností morfologických metod v této kapitole vyžaduje pokročilejší statistické metody. V této kapitole jsou proto některé postupy pouze naznačeny. Podrobně jsou příslušné techniky vysvětleny až v kapitole následující.

Než se pustíme do všech těchto úkolů, musíme si stanovit, co bychom od morfologických metod chtěli a očekávali.

Porovnatelnost. Metody musí umět porovnat rastrový obrázek z elektronového mikroskopu a výsledky modelů, které mohou mít vektorový charakter.

Diskriminabilita. Metody by měly být schopny rozlišit, jestli dva různé obrázky pocházejí z jednoho a téhož experimentu. Metody by měly být schopny citlivě zachycovat změnu charakteru tenké vrstvy v průběhu její tvorby.

Robustnost. Metody by neměly být příliš citlivé na malé či lokální změny v obrázku.

Invariantnost. Metody by neměly být citlivé například na změnu měřítka obrázku, rotaci obrázku či jeho zrcadlení.

Efektivnost. Metody by neměly být extrémně výpočetně náročné.

Nezávislost. Rádi bychom dostali sadu charakteristik, kde každá charakteristika by popisovala jiný rys obrázku. Charakteristiky ani příznaky by tedy neměly být vzájemně korelovány.

Úplnost. Pokud bychom dokázali pomocí morfologických metod plně postihnout informaci skrytou v obrázku, měli bychom být schopni z charakteristik a příznaků obrázek zpětně zrekonstruovat.

K posuzování těchto vlastností je vhodné využít statistických metod a pojmů z teorie pravděpodobnosti. Právě metody využívající statistické zpracování dat nám mohou zajistit robustnost, invariantnost a nezávislost. V této kapitole uvedeme pouze základní pojmy teorie pravděpodobnosti, které jsou nutné k matematickému popisu morfologických metod. (Statistické metody potřebné ke spočtení charakteristik a příznaků budou náplní kapitoly 4.) Následovat bude popis několika konkrétních morfologických metod a jejich otestování na modelových i experimentálních datech.

3.1 Základy teorie pravděpodobnosti

Představme si, že chceme analyzovat vlastnosti celé tenké vrstvy. K tomu bychom potřebovali vyfotografovat každou její část, obrázky navzájem pospojovat a výslednou obří fotografii analyzovat. Tento přístup by byl velice pracný a časově i výpočetně náročný. V praxi míváme k dispozici jen několik málo snímků. Čím méně jich je, tím méně naše údaje vypovídají o celku. Přesto nám však o celku mohou dát dobrou představu.

Na vlastnosti celé tenké vrstvy se můžeme dívat jako na náhodné jevy s neznámou hustotou pravděpodobnosti. Hodnoty získané z jednotlivých fotografií pak můžeme brát jako navzájem nezávislé realizace z daného pravděpodobnostního rozdělení. Naším cílem bude z naměřených dat co nejlépe odhadnout skutečné rozdělení (tedy hustotu pravděpodobnosti či distribuční funkci rozdělení).

Nyní stručně ke značení. Jednotlivé náhodné jevy budeme značit X, resp. Y, Z. Hustotu pravděpodobnosti jevu X budeme označovat jako $\rho_X(x)$, distribuční funkci jako $F_X(x)$. Jednotlivé realizace jevu budeme označovat x_i.

Pro úplnost jen připomeňme, že distribuční funkce je neklesající, zprava spojitá funkce splňující $F_X(x) = P(x_i \leq x)$. (Hodnota distribuční funkce jevu X v bodě x se rovná pravděpodobnosti, že při realizaci jevu dostaneme hodnotu nižší nebo rovnou x.) Hustotu pravděpodobnosti definujeme jako $\rho_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x}$. Hustota pravděpodobnosti pak bude nezáporná funkce splňující $\int_{-\infty}^{\infty} \rho_X(x)\mathrm{d}x = 1$ a $P(a \leq x_i < b) = \int_a^b \rho_X(y)\mathrm{d}y$.

Pokud půjde o odhady, budeme k příslušným veličinám přidávat stříšku: $\hat{\rho}_X(x)$, $\hat{F}_X(x)$, k optimálním odhadům budeme přidávat ještě hvězdičku $\hat{\rho}^*_X(x)$, $\hat{F}^*_X(x)$.

Empirická distribuční funkce je jednoduchým odhadem distribuční funkce. Pro soubor hodnot $\{x_i\}_{i=1}^N$ definujeme empirickou distribuční funkci předpisem $\hat{F}_N(x) = \frac{1}{N}\sum_{i=1}^N I(x \geq x_i)$, $I(A)$ značí indikátor množiny A. Výsledné rozdělení je diskrétní a plně závisí na hodnotách $\{x_i\}_{i=1}^N$. Při náhodném výběru z tohoto rozdělení je každá z hodnot x_i realizována s pravděpodobností N^{-1}.

3.2 Jednotlivé morfologické metody

Vlastnosti, které nám mohou morfologické metody poskytnout, můžeme zjednodušeně rozdělit do následujících skupin: [3]

- Integrální informace - informace o obrázku jako celku
- Informace o jednotlivých objektech
- Informace o rozmístění objektů na ploše/v prostoru
- Informace získané z obrázků zachovávajících stupně šedi

V této práci budeme používat předzpracované obrázky. Objekty budou značeny černou barvou a pozadí bude bílé. Budeme se tedy zabývat jen prvními třemi skupinami. U třetí skupiny budeme uvažovat jen rozmístění v ploše a naznačíme, jak by se daly tyto metody zobecnit pro trojrozměrná data.

3.2.1 Matematický popis experimentálních dat

Abychom mohli metody jednoznačně popsat, musíme umět matematicky popsat některé základní pojmy, jako jsou obrázek, objekt či pixel.

Obrázkem (obecně šedotónovým) budeme myslet dvojici $\{\Omega, \mathcal{I}\}$, kde $\Omega \subset \mathbb{R}^2$ a $\mathcal{I} : \Omega \to [0, 1]$. Množina Ω je nosič obrázku, zpravidla bude čtvercová či obdélníková.

Binarizovaný obrázek pro nás bude obrázek, kde $\mathcal{I} : \Omega \to \{0, 1\}$. V této kapitole budeme dále uvažovat jen obrázky binarizované. Hodnota 0 bude značit bílou barvu, hodnota 1 černou. Pixel budeme považovat za jednotkový čtverec. Pro fotografii o rozměrech $m \times n$ pixelů volíme $\Omega = [0, m] \times [0, n]$.

Diskrétní obrázek pro nás bude takový obrázek, jehož nosič splňuje podmínku $\Omega = [0, m] \times [0, n]$, $m, n \in \mathbb{N}$ a $\mathcal{I}(x) = i_{k,l} \in [0, 1]$ pro $x \in [k, k + 1] \times [l, l + 1]$, $l \in \{0, 1, \ldots, m - 1\}$, $k \in \{0, 1, \ldots, n - 1\}$.

Množinu objektů $\{p_i\}$ definujme pro binarizovaný obrázek jako $\mathcal{P} = \{x \,|\, \mathcal{I}(x) = 1\}$. Jednotlivé objekty $p_i \in \Omega$ pak jsou maximální podmnožiny $\mathcal{P}$ takové, že průnik jejich vnitřků je prázdná množina.

Hraniční objekt pro nás bude takový objekt p_i, pro který platí, že $\overline{p_i} \cap \partial\Omega \neq \emptyset$. Znak ∂ zde značí hranici množiny vzhledem k eukleidovské metrice v $\mathbb{R}^2$, znak $\overline{p_i}$ uzávěr množiny p_i vzhledem k téže metrice.

Dále zavedeme tři různé funkce, které měří vzdálenost objektů p_i, p_j. První z nich definujeme jako $dist_1(p_i, p_j) := inf_{x \in p_i, y \in p_j}\{||x - y||_e\}$, kde $||.||_e$ je eukleidovská norma v $\mathbb{R}^2$. Druhou funkci zavedeme jen pro objekty z diskrétního obrázku $dist_2(p_i, p_j) := dist_1(s_i, s_j) - 1$, kde s_i je množina středů pixelů objektu p_i. Tato definice má za cíl to, aby objekty, které se dotýkají jen rohem, měly nenulovou vzdálenost a naopak aby pixely se společnou hranou patřily do jednoho objektu a měly tedy nulovou vzájemnou vzdálenost. To funkce $dist_1$ nezajišťuje, ale funkce $dist_2$ ano. Třetí funkce ($dist_3$) měří vzdálenost těžišť objektů pomocí $dist_1$.

Velikost (obsah) objektu definujeme jako $|p_i| = \int_{p_i} x \mathrm{d}x$.

Nyní, když už umíme modelová i experimentální data dostatečně přesně popsat, můžeme přistoupit k samotným morfologickým metodám.

3.2.2 Integrální informace

Integrální informace nám dávají informaci o obrázku jako celku. Dvě základní morfologické metody spadající do této kategorie jsou **počet objektů** a **stupeň zaplnění** (někdy také stupeň pokrytí).

Díky předchozím definicím můžeme za počet objektů formálně brát počet prvků množiny $\mathcal{P}$. Menším problémem jsou hraniční objekty, u kterých můžeme předpokládat, že nejsou v Ω obsaženy celé, nemáme o nich tedy úplnou informaci. Pokud budeme znát nebo dokážeme odhadnout tvar objektu, budeme započítávat do počtu objektů jen poměr obsahu objektu, který je uvnitř, vůči celé velikosti objektu. Pokud tvar objektů neznáme, budeme hraniční objekt započítávat jako polovinu objektu.

Stupeň zaplnění nám určuje, jakou část obrázku objekty zabírají. Zde je implementace přímočará. U analyticky zadaných objektů sečteme jejich obsah a podělíme velikostí plochy, v níž jsou umístěné. U diskrétních obrázků stačí takto počítat pixely. Obecně můžeme stupeň zaplnění vyjádřit jako $|\mathcal{P}|/|\Omega|$.

Integrální informace jsou univerzální a můžeme je využít, kdykoli umíme identifikovat jednotlivé objekty a změřit jejich velikost. Lze je jednoduše zobecnit i pro trojrozměrná data.

3.2.3 Popis jednotlivých objektů

Jednotlivé objekty jsou definované svou velikostí a tvarem. Naším cílem je získat představu o těchto vlastnostech, a to ideálně pro všechny objekty (ostrůvky) v celém experimentu. Fotografie z elektronového mikroskopu nám poskytuje informace o malém zlomku těchto objektů. Z těchto informací pak můžeme celkové rozdělené velikostí a tvarů odhadnout.

U velikostí objektů je situace celkem jednoduchá. Velikost objektů jsme definovali jako $\int_{p_i} x\mathrm{d}x$, jde tedy o číselné veličiny. Velikosti objektů na obrázku můžeme brát jako náhodná čísla vygenerovaná z rozdělení s neznámou hustotou pravděpodobností. Naším cílem je tuto hustotu co nejlépe odhadnout. Odhady hustoty pravděpodobnosti se zabývá sekce 4.2.

Pokud chceme postihnout tvar objektů, dostáváme se do složité situace. Jakýkoli příznak bude nutně obsahovat jen část informace o tvaru objektu. Při volbě příznaků tedy velmi záleží na tom, jaký aspekt tvaru chceme postihnout. Během tvorby tenkých vrstev jsou průměty objektů zpočátku kruhové a teprve v pozdějších fázích se jejich tvar začíná diferencovat. Přirozeným požadavkem je tedy měřit „míru kruhovosti“, k čemuž nám může sloužit například tzv. tvarový faktor (anglicky form factor). Ten je definován pro jeden objekt jako normovaný poměr jeho obsahu a druhé mocniny jeho obvodu: $FF(p_i) = 4\pi\frac{|p_i|}{|\partial p_i|^2}$ [3]. Tvarový faktor může nabývat hodnot mezi nulou a jedničkou. Pro kruhový objekt získáváme tvarový faktor roven jedné. Čím více se tvar odlišuje od kruhového, tím nižší má tvarový faktor. Všimněme si, že hodnota tvarového faktoru nezávisí na velikosti objektu. Příznak je invariantní vůči škálování (změně měřítka).

Nevýhodou tvarového faktoru je problematický výpočet obvodu u diskrétních obrázků. Proto zavádíme ještě jednu veličinu, která se výpočtu obvodu objektu vyhýbá

- acirkularitu. Ta vychází z porovnání objektu p_i s kruhovým objektem o stejném obsahu p_i^{ref}. Pokud tyto dva objekty překryjeme (střed kruhového objektu položíme do těžiště původního objektu), můžeme porovnat, jak se původní objekt od kruhového liší. Acirkularitu definujme jako podíl počtu pixelů tvořících p_i majících s p_i^{ref} nulový průnik (tedy poměr pixelů, které p_i^{ref} nezakrývá ani trochu ku počtu pixelů celého objektu). Pro kruhové objekty tedy bude acirkularita nulová, s rostoucí nepravidelností tvaru acirkularita roste k jedné. Pro analyticky zadané objekty můžeme acirkularitu počítat pomocí integrálů.

Oba popsané příznaky lze použít, kdykoli známe tvar objektů, lze je rovněž modifikovat pro trojrozměrná data. Musíme však mít na paměti, že oba představené příznaky zachycují jen jeden aspekt tvaru objektu. Pro pozdější fáze růstu tenkých vrstev, kdy se tvary objektů začínají diferencovat a začíná vznikat nespojitá struktura, nemusí být takovýto popis postačující.

3.2.4 Radiální distribuční funkce

Nyní se přesuneme ke třetí skupině metod — k metodám, které zkoumají rozmístění objektů na ploše. Příkladem takovéto veličiny je tzv. radiální distribuční funkce. Poskytuje informace o vzájemné vzdálenosti objektů následujícím způsobem: zjišťuje, jaká je pravděpodobnost, že dva náhodně zvolené objekty jsou od sebe vzdálené méně než r, pokud platí, že jsou od sebe vzdálené méně než r_{max}.

Výklad v této sekci bude probíhat opačně než v sekci předchozí. Nejprve rozebereme definici jevu samotného a teprve z ní odvodíme, jak získat odhady jeho hustoty pravděpodobnosti z měření na obrázku. Začněme s definicí hustoty pravděpodobnosti:

$$\rho_{RDF}(r) = \frac{C}{r} \frac{\mathrm{d}P(dist(p_i, p_j) \leq r \,|\, p_i \neq p_j, dist(p_i, p_j) \leq r_{max})}{\mathrm{d}r}\,, \tag{3.1}$$

kde $dist(i, j)$ je zatím nespecifikovaná verze funkce pro měření vzdálenosti i-tého a j-tého objektu.

Pokud bude rozložení objektů blízké homogennímu, poroste $P(dist(p_i, p_j) \leq r)$ se vzdáleností r řádově kvadraticky a $\frac{\mathrm{d}P(dist(p_i,p_j)\leq r)}{\mathrm{d}r}$ řádově lineárně. Škálování pomocí členu $1/r$ nám zajistí, že $\rho_{RDF}(r)$ bude blízké rovnoměrnému rozdělení, což nám pomůže při odhadech této hustoty. Konstanta C pak zajišťuje normování hustoty pravděpodobnosti tak, aby $\int_0^{r_{max}} \rho_{RDF}(r)\mathrm{d}r = 1$.

Co si pod vzorcem (3.1) představit?

Zvolme si náhodně objekt a okolo něho vytvořme soustředné kružnice o poloměrech $i \cdot \Delta r$ pro $i = 1 \ldots \lceil r_{max}/\Delta r \rceil$. (Správně by v tomto místě měl následovat limitní přechod $\Delta r \to 0$, ten však nyní z didaktických důvodů vynecháme.) Pro každý kruh nyní spočítáme, kolik objektů uvnitř něho leží. Tím dostáváme (až na konstantu) odhad členu $P(dist(p_i, p_j) \leq r \,|\, p_i \neq p_j, dist(p_i, p_j) \leq r_{max})$.

Derivaci $\frac{\mathrm{d}P(dist(p_i,p_j)\leq r \,|\, p_i\neq p_j, dist(p_i,p_j)\leq r_{max})}{\mathrm{d}r}$ můžeme aproximovat diferencí. Výraz pak můžeme vyčíslit tak, že místo počtu objektů v kruhu o poloměru $i \cdot \Delta r$ nyní počítáme počet objektů v mezikruží s poloměry $i \cdot \Delta r$ a $(i + 1) \cdot \Delta r$. Nakonec pro každé mezikruží podělíme počet objektů jeho vnitřním poloměrem (pokud bychom prováděli

vynechaný limitní přechod, nezáviselo by na tom, jestli bereme vnitřní či vnější poloměr). Pokud tento proces zopakujeme pro všechny objekty, dostaneme (po částech konstantní) odhad veličiny ρ_{RDF}. Takto je popsán výpočet radiální distribuční funkce například v [3].

Je zřejmé, že přesnost odhadu závisí na velikosti Δr. Příliš velké Δr může zakrýt změny průběhu veličiny ρ_{RDF}, pokud zvolíme Δr naopak příliš malé, průběh překryjí náhodné fluktuace. Další nepřesnost výše uvedené aproximace spočívá v tom, že všechny objekty ležící v jednom mezikruží jsou brány se stejnou váhou. V definici veličiny (3.1) je každá dvojice objektů brána s váhou rovnou převrácené hodnotě jejich vzdálenosti. Oproti tomu v popsané aproximaci je jako váha brána převrácená hodnota vnitřního poloměru mezikruží. Tato nepřesnost se dá odstranit tak, že nebudeme váhovat celkový počet objektů v mezikruží, ale už od začátku budeme každý objekt započítávat ne jedenkrát, nýbrž $1/r$-krát.

Dalším problémem může být omezená velikost obrázku. Pokud bychom jako centrální objekt, kolem kterého uvažujeme kružnice, zvolili objekt příliš blízký okraji obrázku, některá mezikruží by přesáhla okraj obrázku. Abychom omezili vliv okrajů oblasti, budeme do vztahu (3.1) za p_i volit jen objekty vzdálené od okraje oblasti více než r_{max}.

Pro bodové objekty je definice (3.1) jednoznačná. Pro diskrétní obrázky je potřeba specifikovat, jakou verzi funkce *dist* uvažujeme. První možností je brát vzdálenost objektů jako vzdálenost jejich težišť a použít k měření vzdáleností funkci $dist_1$. Takto definovanou veličinu budeme označovat jako ρ_{RDF_1}.

Druhou variantou je měřit vzdálenost mezi okraji objektů pomocí $dist_2$. V tomto případě však není rozumné škálování pomocí členu $1/r$, což se projeví zejména pro malé hodnoty r. Místo obecného škálování zavádíme škálování pro každé měření zvlášť. Pokud máme množinu vzdáleností změřenou pomocí $dist_2$, každé měření budeme brát s váhou nepřímo úměrnou vzdálenosti jejich těžišť (tedy nepřímo úměrnou jejich vzdálenosti měřené funkcí $dist_3$). Takto definovanou veličinu budeme označovat jako ρ_{RDF_2}. V kompaktním tvaru můžeme tuto veličinu zapsat následovně:

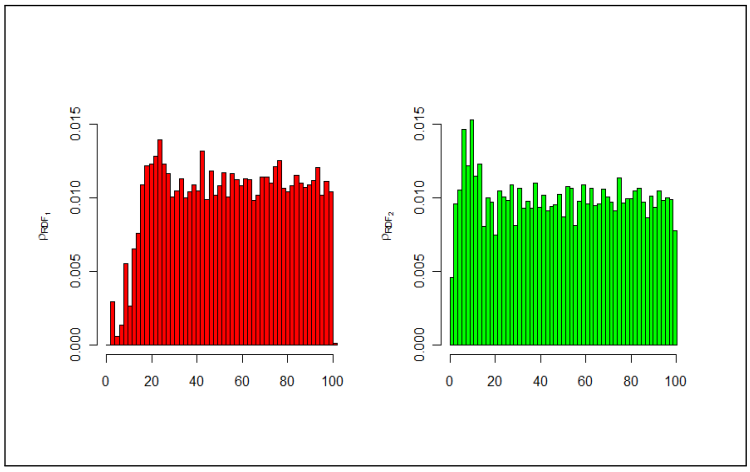
$$\rho_{RDF}(r) = \frac{C}{r} \frac{\mathrm{d}P(dist_3(p_i,p_j) \leq r \,|\, p_i \neq p_j, dist_2(p_i,p_j) \leq r_{max})}{\mathrm{d}(r + dist_2(p_i,p_j) - dist_3(p_i,p_j))} \,. \tag{3.2}$$

Poslední variantu radiální distribuční funkce budeme používat pro kruhové objekty. Postup výpočtu je analogický variantě druhé, jedinou změnou je použití funkce $dist_1$ místo $dist_2$ pro měření vzdáleností objektů. Váhování pomocí funkce $dist_3$ zde zůstává zachováno. Tuto variantu budeme označovat jako ρ_{RDF_3}.

V praxi nelze najít ρ_{RDF} přesně, jako charakteristiku obrázku používáme její odhad. Na obrázku 3.1 vidíme aplikaci námi definovaných radiálních distribučních funkcí na obrázek 1.3. Pro odhad hustot pravděpodobností byl použit histogram, viz část 4.2.3.

Vidíme, že v obou případech má hustota pravděpodobnosti jedno výrazné maximum, které odpovídá nejčastější vzdálenosti nejbližších objektů. Čím je struktura objektů pravidelnější, tím bude toto maximum vyšší. U pravidelných struktur by se navíc objevilo i celkové zvlnění hustoty pravděpodobnosti a objevila by se i další maxima.

Vidíme, že modifikace váhování první maximum zvýraznila a omezila tak vliv rozdílných velikostí objektů. Pro zpracování experimentálních dat tedy doporučujeme



Obrázek 3.1: Porovnání dvou variant radiální distribuční funkce ρ_{RDF_1} a ρ_{RDF_2}

používat tuto variantu metody.

Jako příznak můžeme u radiální distribuční funkce použít například podíl velikosti prvního maxima a prvního minima. Odhady těchto hodnot se zabývá část 4.2.4.

3.2.5 Voronoiovo dláždění a Delaunayova triangulace

Další metody zkoumající vzájemnou pozici objektů jsou **Voronoiovo dláždění**[1] a **Delaunayova triangulace**. Dláždění je způsob, jak rozdělit oblast Ω na konečné množství podoblastí. Triangulace je druh dláždění, kde výsledné podoblasti budou mít tvar trojúhelníku (simplexu pro $\Omega \subset \mathbb{R}^d, d > 2$). Pokud je dláždění vhodně spjato s objekty na ploše, lze jej použít k popisu vzájemného postavení objektů v Ω. Obě metody se ve své původní formě používají pro bodové objekty. Lze je však zobecnit i pro plošné či prostorové objekty. I když lze metody definovat pro libovolné $d \in \mathbb{N}$, zůstaneme pro jednoduchost u rovinné verze $d = 2$.

Voronoiovo dláždění můžeme definovat jako množinu dlaždic (tzv. **Voronoiových buněk**) $\mathcal{V}_i \subset \Omega$, jež jsou definovány jako

$$\mathcal{V}_i = \{x \in \Omega \,|\, dist(x, p_i) \leq dist(x, p_j),\, \forall j \neq i\}, \tag{3.3}$$

kde $dist(x, p_i)$ značí vzdálenost bodu x od objektu p_i. Každá dlaždice je tedy množinou bodů, které mají k příslušnému objektu nejblíže.

Pokud budou objekty bodové, množiny bodů, které sdílí více než jedna dlaždice – tzv. hrany dlaždic, budou tvořeny úsečkami. Pokud budou objekty kruhové, hrany Voronoiových buněk budou tvořeny kuželosečkami. Pokud objekty nejsou kruhové, nelze obecně hrany Voronoiových buněk vyjádřit analyticky. Objekty, jejichž Voronoiovy buňky sdílí společnou hranu, budeme nazývat **přirozenými sousedy**.

Body, které sdílí více než dvě Voronoiovy buňky, budeme nazývat vrcholy Voronoiových buněk. Ve většině případů náleží vrcholy právě třem dlaždicím. Může však nastat

[1] V angličtině se používají termíny Voronoi diagram, Voronoi tessellation, Voronoi decomposition či Dirichlet tessellation

situace, kdy některý z vrcholů náleží čtyřem nebo více dlaždicím. Takové dláždění pak budeme nazývat degenerovaným.

Pro Ω konvexní jsou Voronoiovy buňky souvislé množiny. Pokud je navíc Ω polygonální a objekty jsou bodové, mají Voronoiovy buňky tvar polygonu. Tímto případem se nyní budeme podrobněji zabývat. Přidejme ještě dva předpoklady na množinu objektů:

Nekolinearita. Body množiny $\mathcal{P}$ neleží na jedné přímce.

Necirkularita. Neexistuje kružnice taková, že by čtyři nebo více objektů leželo na této kružnici a všechny ostatní objekty byly vně kružnice.[2]

Pokud platí všechny tyto předpoklady a množina $\mathcal{P}$ obsahuje alespoň tři body, propojením bodů, které jsou si navzájem přirozenými sousedy, získáme dláždění konvexního obalu množiny objektů. Toto dláždění se nazývá Delaunayovou triangulací $\mathcal{D} = \{\mathcal{T}_i\}$. Trojúhelník triangulace $\mathcal{T}_i$ je množina bodů splňující

$$\mathcal{T}_i = \{x \mid x = \sum_{j=1}^{3} \lambda_j x_{ij}, \sum_{i=1}^{3} \lambda_j = 1, \lambda_j \geq 0\}, \tag{3.4}$$

přičemž x_{ij} jsou body, jejichž Voronoiovy dlaždice mají společný vrchol. Tyto vrcholy tvoří středy kružnice opsané trojúhelníkům triangulace.[3]

Další důležitou vlastností těchto dláždění je jejich lokálnost. Změna pozice jednoho objektu ovlivní jen omezený počet dlaždic. Další vlastnosti je možné najít v [7] stejně jako důkazy zde uvedených tvrzení.

Pro Voronoiovo dláždění a Delaunayovu triangulaci existuje hned celá řada příznaků, které mohou postihnout pravidelnost struktury objektů. Můžeme zkoumat velikost dlaždic, délku hranice jednotlivých dlaždic nebo například počet sousedů dlaždice. Tyto údaje můžeme také libovolně kombinovat.

Příkladem příznaku může být veličina $\frac{\sqrt{\operatorname{Var}|\mathcal{V}_i|}}{\operatorname{E}|\mathcal{V}_i|}$, která měří poměr odmocniny rozptylu velikostí Voronoiových buněk ku střední hodnotě velikosti buněk. Statistický charakter spolu s lokálností Voronoiova dláždění nám zajistí dostatečnou robustnost. Bezrozměrnost příznaku nám dává invarianci příznaku vůči změně měřítka.

Definujme nyní několik takových příznaků a sledujme jejich průběh pro výsledky modelu hard-disk s různým poměrem zaplnění.

Chování příznaků je zobrazeno na obrázcích 3.2 a 3.3. Analyzována byla data z modelu hard-disk pro parametry $\Omega = [0;1] \times [0;1]$, $R = 0{,}005$, $P = 0{,}03 : 1 : 0{,}01$ [4] a $D = 0{,}01$. Vynesené chyby odpovídají bootstrapovým odhadům směrodatné odchylky pro $B = 1\,000$.[5]

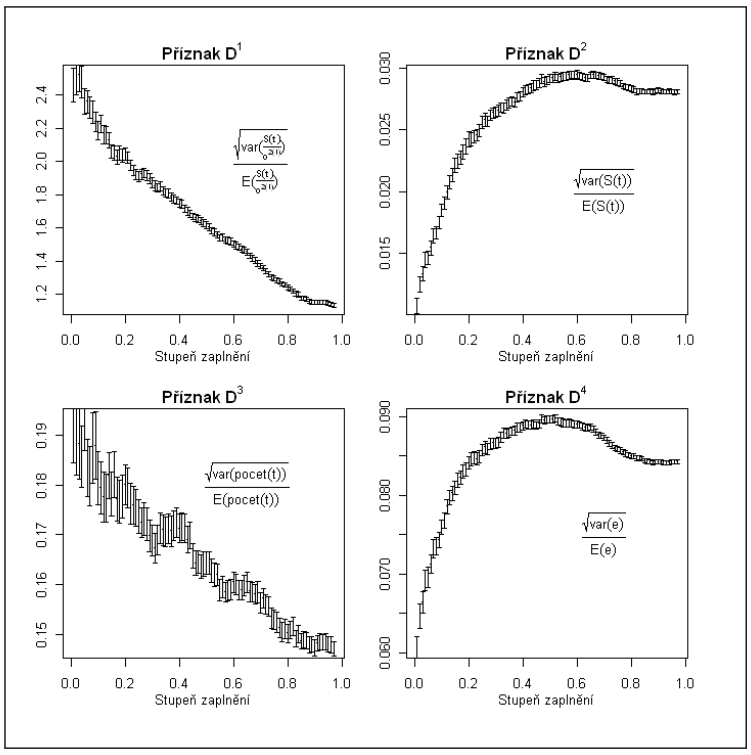
Jak je vidět z grafů, příznaky můžeme podle jejich průběhu rozdělit na dvě skupiny. Zatímco příznaky D^1, D^3, V^1 a V^3 s rostoucím stupněm zaplnění klesají, příznaky D^2, D^4, V^2 a V^4 nejprve rostou a později stagnují či dokonce mírně klesají.

[2] Necirkularita je ekvivalentní tomu, že Voronoiovo dláždění není degenerované.

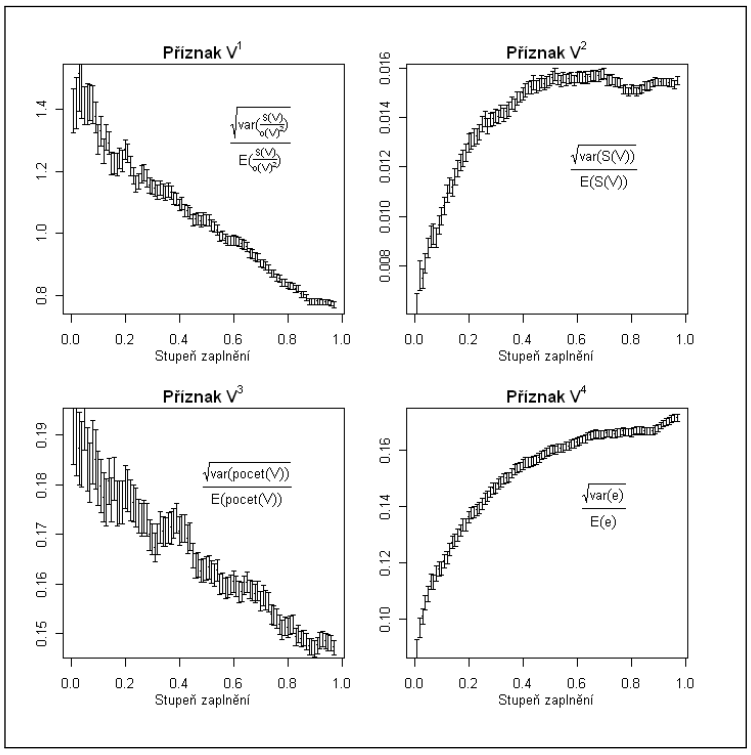
[3] V případě, že by nebyla splněna podmínka necirkularity, by body x_{ij} byly pro nějaké i alespoň čtyři a definice $\mathcal{T}_i$ by musela být mírně poupravena.

[4] Značení $a : b : c$ značí posloupnost čísel začínající a, jdoucí po krocích c až do b.

[5] Vysvětlení principu bootstrapových odhadů se věnuje část 4.3.



Obrázek 3.2: Závislost příznaků D^1-D^4 na stupni zaplnění u hard-disk modelu



Obrázek 3.3: Závislost příznaků V^1-V^4 na stupni zaplnění u hard-disk modelu

Označení	Vzorec	Slovní popis klíčové veličiny
V^1	$\frac{\sqrt{\mathrm{Var}\,\frac{\lvert\mathcal{V}_i\rvert}{\lvert\partial\mathcal{V}_i\rvert^2}}}{\mathrm{E}\,\frac{\lvert\mathcal{V}_i\rvert}{\lvert\partial\mathcal{V}_i\rvert^2}}$	Podíl obsahu ku druhé mocnině obvodu Voronoiových dlaždic
V^2	$\frac{\sqrt{\mathrm{Var}\,\lvert\mathcal{V}_i\rvert}}{\mathrm{E}\,\lvert\mathcal{V}_i\rvert}$	Obsah Voronoiových dlaždic
V^3	$\frac{\sqrt{\mathrm{Var}(\text{počet sousedů } \mathcal{V}_i)}}{\mathrm{E}(\text{počet sousedů } \mathcal{V}_i)}$	Počet sousedních buněk
V^4	$\frac{\sqrt{\mathrm{Var}(\text{délky hran } \mathcal{V}_i)}}{\mathrm{E}(\text{délky hran } \mathcal{V}_i)}$	Délky hran Voronoiových dlaždic
D^1	$\frac{\sqrt{\mathrm{Var}\,\frac{\lvert\mathcal{T}_i\rvert}{\lvert\partial\mathcal{T}_i\rvert^2}}}{\mathrm{E}\,\frac{\lvert\mathcal{T}_i\rvert}{\lvert\partial\mathcal{T}_i\rvert^2}}$	Podíl obsahu ku druhé mocnině obvodu trojúhelníků Delaunayovy triangulace
D^2	$\frac{\sqrt{\mathrm{Var}\,\lvert\mathcal{T}_i\rvert}}{\mathrm{E}\,\lvert\mathcal{T}_i\rvert}$	Obsah obsahu trojúhelníků Delaunayovy triangulace
D^3	$\frac{\sqrt{\mathrm{Var}(\text{počet trojúhelníků obsahujících } p_i)}}{\mathrm{E}(\text{počet trojúhelníků obsahujících } p_i)}$	Počet trojúhelníků Delaunayovy triangulace stýkajících se v jednom bodě
D^4	$\frac{\sqrt{\mathrm{Var}(\text{délky hran } \mathcal{T}_i)}}{\mathrm{E}(\text{délky hran } \mathcal{T}_i)}$	Délky hran trojúhelníků Delaunayovy triangulace

Tabulka 3.1: Přehled příznaků založených na Voronoiově dláždění a Delaunayově triangulaci

Příznaky v první skupině jsou navzájem velice silně korelovány, jak nám ukazuje tabulka 3.2. Příznaky z druhé skupiny jsou navzájem korelovány méně. Obě skupiny jsou pak vůči sobě korelovány záporně. Analýza hlavních komponent ukazuje, že pokud vhodně vybereme z osmi příznaků dvě jejich lineární kombinace, dokážeme popsat téměř sto procent informací získaných z příznaků (první hlavní komponenta popisuje 88,8 procenta variance dat, druhá pak 10,7 procenta).

	D^1	D^2	D^3	D^4	V^1	V^2	V^3	V^4
D^1	1,000	-0,835	0,986	-0,618	0,997	-0,865	0,986	-0,951
D^2	-0,835	1,000	-0,805	0,945	-0,819	0,996	-0,800	0,958
D^3	0,986	-0,805	1,000	-0,584	0,984	-0,835	0,999	-0,929
D^4	-0,618	0,945	-0,584	1,000	-0,596	0,923	-0,577	0,821
V^1	0,997	-0,819	0,984	-0,596	1,000	-0,850	0,984	-0,938
V^2	-0,865	0,996	-0,835	0,923	-0,850	1,000	-0,830	0,973
V^3	0,986	-0,800	0,999	-0,577	0,984	-0,830	1,000	-0,925
V^4	-0,951	0,958	-0,929	0,821	-0,938	0,973	-0,925	1,000

Tabulka 3.2: Korelační matice příznaků D^1-D^4, V^1-V^4 aplikovaných na hard-disk model

Pokusme se nyní interpretovat chování příznaků. Do první skupiny patří příznaky odpovídající tvarovému faktoru dlaždic a příznaky založené na počtu sousedů dlaždice.[6] Hodnota těchto příznaků tedy závisí jen na tvaru dlaždic bez ohledu na jejich

[6]Povšimněme si, že příznaky D^3 a V^3 se od sebe téměř neliší. Oba měří tu samou veličinu. Rozdílné hodnoty v korelační matici jsou dány jen jiným počtem zahrnutých dlaždic.

velikost. Oproti tomu příznaky z druhé skupiny závisí jak na tvaru dlaždic, tak i na jejich velikosti.

Podívejme se, jaké hodnoty příznaků dostaneme, pokud metody aplikujeme na obrázek 1.3. Jako pozice objektů v tomto případě budeme brát jejich těžiště. Výsledky jsou uvedené v tabulce 3.3.

D^1	1,855 ± 0,037
D^2	0,045 ± 0,001
D^3	0,173 ± 0,003
D^4	0,112 ± 0,001
V^1	1,176 ± 0,024
V^2	0,026 ± 0,001
V^3	0,173 ± 0,003
V^4	0,195 ± 0,002

Tabulka 3.3: Hodnoty příznaků D^1-D^4, V^1-V^4 aplikovaných na obrázek 1.3

Vidíme, že pro příznaky z první skupiny odpovídají hodnoty stupni zaplnění cca 0,3. Oproti tomu hodnoty příznaků druhé skupiny přesáhly maximum hodnot u hard-disk modelu. Příčinou tohoto jevu je různý poloměr objektů na obrázku. Situaci si můžeme představit tak, že každému objektu (v tomto případě těžišti objektu) náleží různý difúzní parametr, který úzce souvisí s velikostí objektu. Čím má objekt větší poloměr, tím větší difúzní zóna mu přísluší. V námi použité verzi hard-disk modelu měly všechny objekty stejnou velikost. Na obrázku se ale velikosti objektů liší, což se pak promítá do větší variability velikosti dlaždic a výsledně i do vyšších hodnot příznaků z druhé skupiny.

Naskýtá se zde otázka, zda by nebylo možné modifikovat příznaky tak, aby byl eliminován nebo alespoň lépe vymezen vliv velikosti objektů na příznaky. Tím se budeme zabývat v následující sekci. Nyní se podívejme, jakým způsobem můžeme hodnotit robustnost a citlivost příznaků.

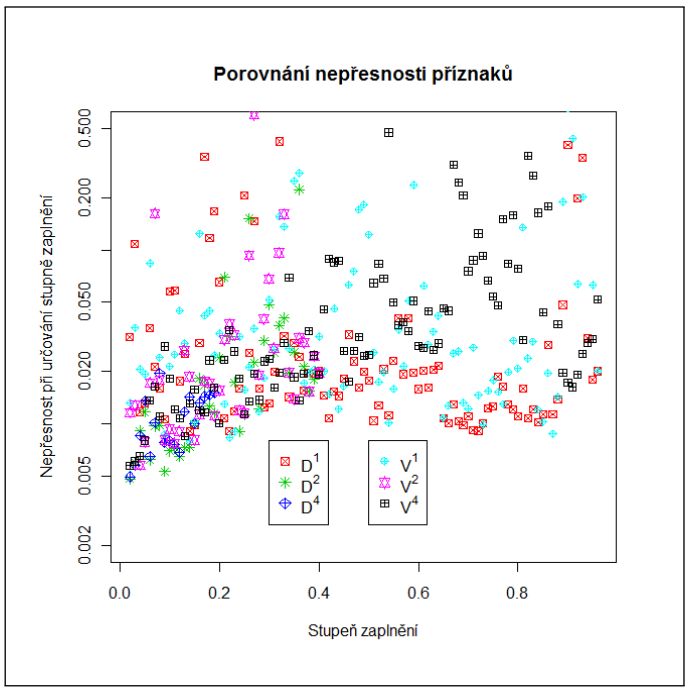
Představme si, že jsme dostali výstup hard-disk modelu se známou velikostí objektů a známým difúzním parametrem. Naším úkolem je co nejlépe odhadnout, jaký byl zvolen poměr zaplnění. Je jasné, že nejlepší odhad bychom dostali zkombinováním všech osmi příznaků. Pokud bychom ale měli zvolit jen jeden z nich, jak bychom to provedli?

Pokud bychom uvažovali o příznaku jako funkci závislé na poměru zaplnění, mohli bychom stupeň zaplnění získat jako hodnotu inverzní funkce pro naměřenou hodnotu příznaku. Pro příznaky D^1 a V^1 by tento přístup měl fungovat, příznaky jsou totiž monotónními funkcemi vzhledem ke stupni zaplnění. Naopak příznaky D^3 a V^3 jsou kvůli častým výkyvům špatně použitelné. Příznaky z druhé skupiny lze použít jen pro dostatečně nízké hodnoty, invertovat můžeme jen cca první pětinu až polovinu funkce. Výjimkou je příznak V^4, který je monotónní na celém intervalu. Zkusme nyní porovnat přesnost inverze příznaků na povolených intervalech.

S touto úlohou silně souvisí dva již zmíněné pojmy – robustnost a citlivost. Jako indikátor robustnosti můžeme brát směrodatnou odchylku (čím méně hodnota příznaku pro srovnatelná data kolísá, tím lépe). Jako indikátor citlivosti zvolme převrácenou

hodnotu derivace příznaku podle stupně zaplnění. (Příznak je tím citlivější, čím více se změnou stupně zaplnění změní svojí hodnotu více.) Pokud chceme invertováním průběhu příznaků co nejpřesněji odhadnout stupeň zaplnění, zvolíme příznak, který bude mít nejmenší poměr směrodatné odchylky ku derivaci. Právě tato veličina nám měří, s jakou nejistotou nám příznak odhadne stupeň zaplnění.

Pro porovnání jsme použili všechny příznaky kromě D^3 a V^3. Derivaci jsme aproximovali centrální diferencí. Výsledné hodnoty jsme vynesli do obrázku 3.4. Hodnoty na ose y ukazují, jak velké chyby při určování stupně zaplnění bychom se s různými příznaky mohli dopustit.



Obrázek 3.4: Odhady nepřesnosti vybraných příznaků v závislosti na stupni zaplnění u hard-disk modelu

Vidíme, že pro malé stupně zaplnění (cca 0 - 0,2) jsou přesnější příznaky D^2, D^4, V^2 a V^4. Pro vyšší stupně zaplnění (cca od stupně zaplnění 0,3) pak převažují příznaky D^1 a V^1.

Zpracování experimentálních dat

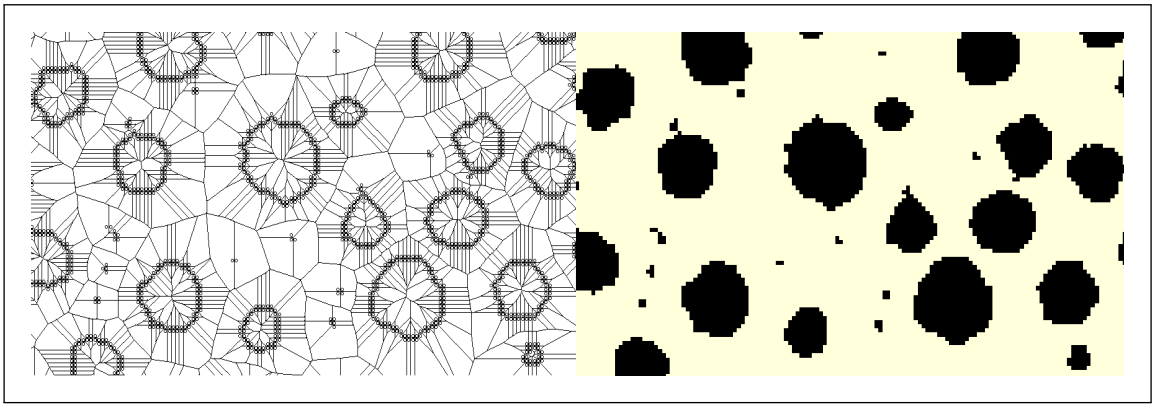
Nyní se vraťme k myšlence omezit vliv velikosti objektů na příznaky vycházející z Voronoiova dláždění a Delaunayovy triangulace. Takovéto příznaky by mohly lépe popisovat prostorové rozložení objektů u experimentálních dat a také u modelů, v nichž se vyskytují objekty různých velikostí. Začněme s Voronoiovým dlážděním.

V předchozí sekci jsme vytvořili Voronoiovo dláždění jen pro těžiště objektů. Takovéto dláždění však neodpovídalo našim intuitivním představám. U některých velkých objektů protínala hranice Voronoiovy dlaždice vnitřek objektu, jemuž náležela. Dále budeme variantu příznaků pracujících jen s těžišti objektů značit dolním indexem t.

Druhou možností je aproximovat objekty kruhy o stejném obsahu a stejném těžišti. Pro kruhové objekty můžeme tvar Voronoiových buněk vyjádřit analyticky. Konstrukce buněk je ale algoritmicky složitá. Navíc zde ztrácíme informace o původním tvaru objektů. Takovouto aproximaci budeme značit dolním indexem $circ$, u Voronoiova dláždění ji však používat nebudeme.

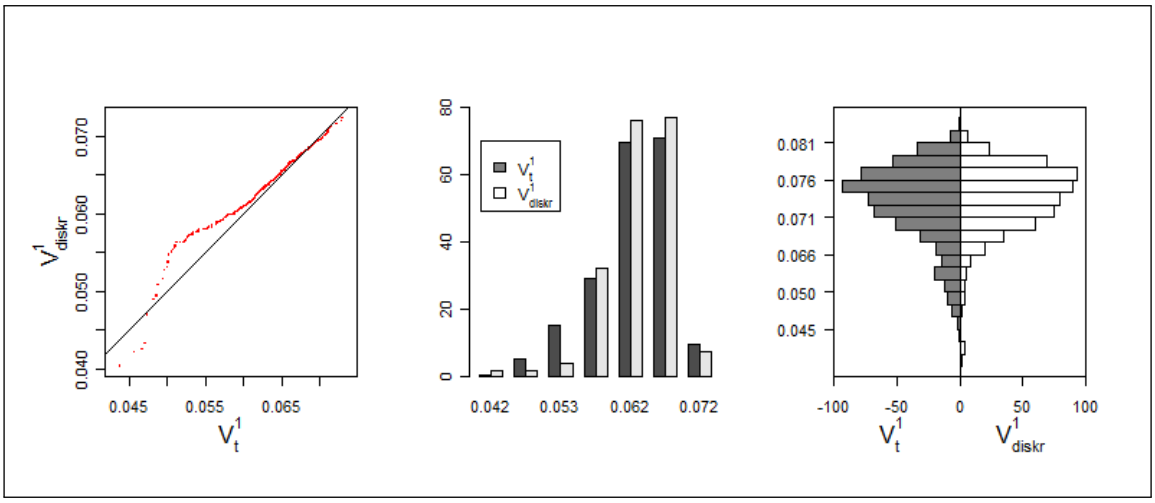
Další možností je využít diskrétního charakteru obrázku a brát Voronoiovy buňky, stejně jako objekty, jako množiny pixelů a pro měření vzdáleností funkci $dist_2$. Zde je zohledněn jak tvar objektů, tak jejich velikost. Problémy této aproximace jsou diskrétní charakter Voronoiových buněk, častější výskyt degenerace (čtyři buňky budou sdílet stejný vrchol) a špatně definované hranice buněk - hranice by tvořily v některých místech celé pixely a v jiných jen jejich okraje.

Metoda, kterou zde budeme analyzovat, řeší všechny zmíněné nevýhody. Funguje následovně. Z každého objektu vybereme hraniční pixely, resp. jejich středy. Tak získáváme množinu bodů, pro které vytvoříme Voronoiovo dláždění. Dlaždice patřící bodům téhož objektu pak sloučíme dohromady. Tím získáme velmi dobrou aproximaci Voronoiova dláždění pro celé objekty. Dlaždice budou mít polygonální tvar a jejich velikost bude součtem velikosti všech buněk, které jsme sloučili. Degeneraci dláždění můžeme zabránit malou perturbací poloh středů hraničních pixelů objektů. Jedinou nevýhodou této metody je velká výpočetní náročnost oproti předchozím možnostem. Takovouto variantu Voronoiova dláždění budeme nazývat „diskrétním Voronoiovým dlážděním“ a příznaky s ní spojené budeme označovat dolním indexem $diskr$. Ukázka diskrétního Voronoiova dláždění je na obrázku 3.5.

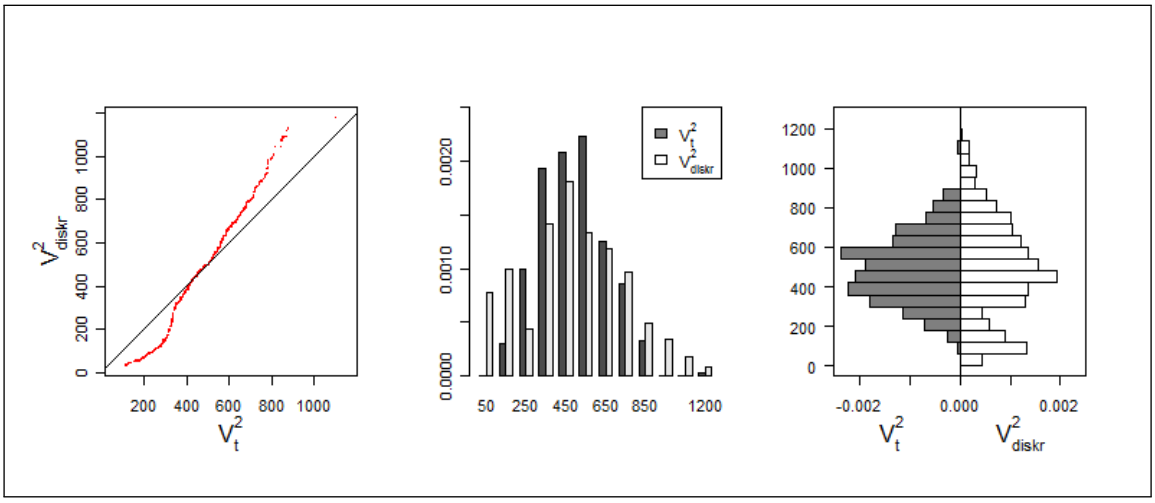


Obrázek 3.5: Vlevo ukázka diskrétního Voronoiova dláždění, vpravo původní obrázek

Pro porovnání jsme spočítali hodnoty příznaků $V_t^1 = 1{,}176 \pm 0{,}024$, $V_t^2 = 0{,}026 \pm 0{,}001$, $V_{diskr}^1 = 1{,}076 \pm 0{,}030$ a $V_{diskr}^2 = 0{,}033 \pm 0{,}001$ pro obrázek 1.3. Vidíme, že citlivější vzhledem ke změně varianty je příznak V^2. Porovnání těchto dvojic aproximací jsou vynesena na obrázcích 3.6 a 3.7 ve formě Q-Q diagramu (viz kapitola 4.1.3), histogramu a stromového grafu. Zatímco varianty příznaku V^1 se od sebe liší jen nepatrně, u příznaku V^2 vidíme diametrální změny. Z obrázků je vidět, že varianta V_{diskr}^2 do jisté míry rozseparovala dlaždice do dvou skupin. Nižší hodnoty velikostí dlaždic můžeme přičíst malým objektům vzniklým sekundární nukleací, jimž varianta $diskr$ zmenšila oproti variantě t buňku.

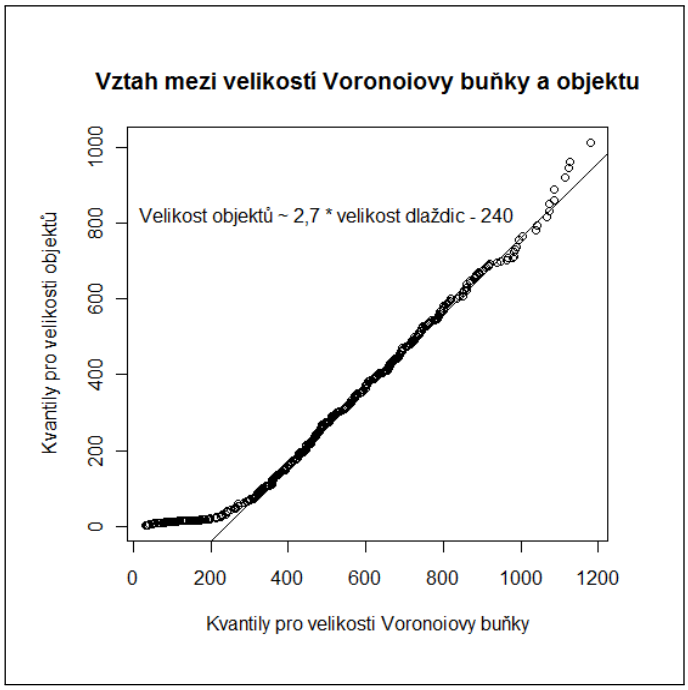


Obrázek 3.6: Porovnání dvou variant příznaku V^1 - příznaků V^1_t a V^2_{diskr} pro obrázek 1.3



Obrázek 3.7: Porovnání dvou variant příznaku V^2 - příznaků V^2_t a V^2_{diskr} pro obrázek 1.3

Navíc podle obrázku 3.8 je velikost dlaždic těsně spjata s velikostí objektů, pokud jsou dostatečně velké. Ve výsledku tedy můžeme pomocí Voronoiova dláždění postihnout jak rozložení objektů, tak i rozdělení jejich velikostí.



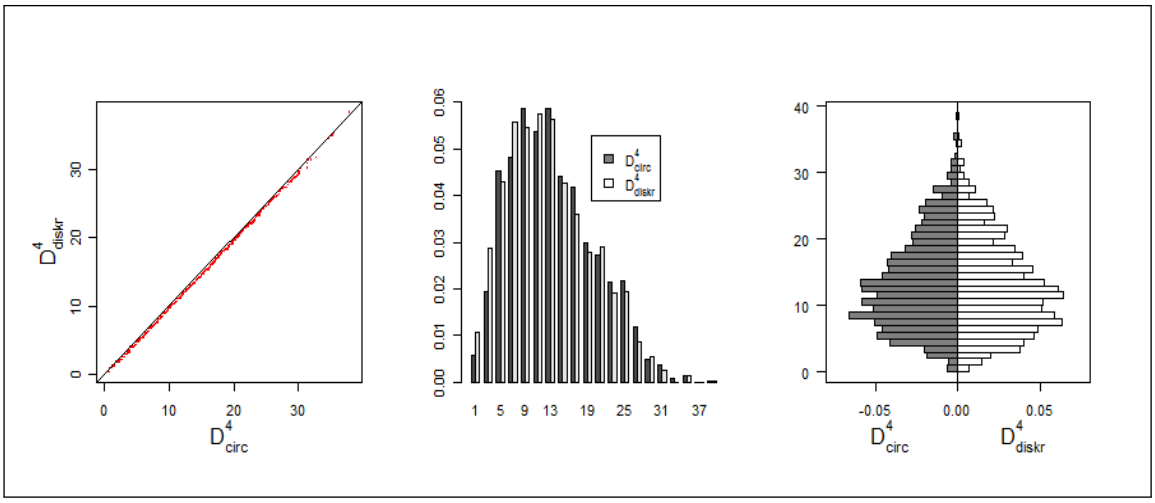
Obrázek 3.8: Q-Q diagram pro velikost Voronoiových buněk a velikosti objektů pro obrázek 1.3

Definice Delaunayovy triangulace je spjata s bodových charakterem objektů těsněji. Pro jednoduchost se zde omezíme jen na zobecnění příznaku D^4, který popisuje rozdělení vzdáleností přirozených sousedů. Z Voronoiova dláždění můžeme určit, které objekty mají sousedící Voronoiovy buňky. Vzdálenosti přirozených sousedů (které pro bodové objekty odpovídají délkám hran trojúhelníku triangulace) pak budeme měřit pomocí funkce $dist_2$. Získáváme tak diskrétní analogii příznaku - D^4_{diskr}.

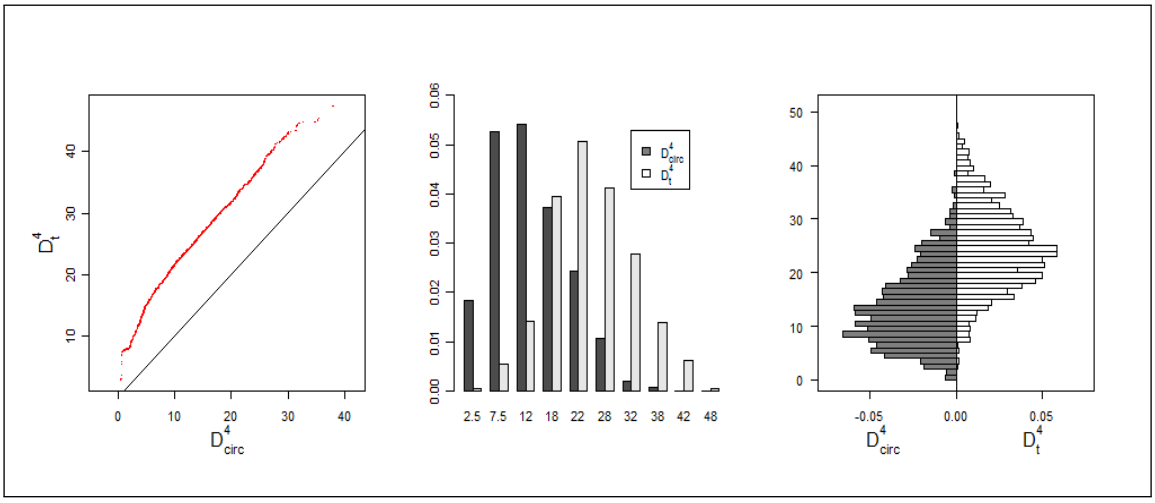
Stejně jako u Voronoiova dláždění zde můžeme zadefinovat i kruhovou a těžišťovou aproximaci D^4_{circ} a D^4_t. Porovnání variant D^4_{diskr} a D^4_{circ} je na obrázku 3.9 ve formě Q-Q diagramu (viz kapitola 4.1.3), histogramu a stromového grafu. Vynesena jsou rozdělení vzdáleností nejbližších sousedů pro obě varianty. Vidíme, že rozdíly v rozdělení nejsou veliké, přesto rozdělení nejsou identická (Kolmogorovův-Smirnovův[7] test dává $p = 0{,}01144$). Hodnoty příznaků jsou $D^4_t = 0{,}112 \pm 0{,}001$, $D^4_{diskr} = 0{,}076 \pm 0{,}001$, $D^4_{circ} = 0{,}195 \pm 0{,}002$.

Porovnání variant D^4_{circ} a D^4_t pak můžeme vidět na obrázku 3.10. Vidíme, že obě distribuce jsou vůči sobě posunuty. Q-Q diagram nám navíc ukazuje, že kromě posunutí se obě distribuce liší i tvarem. Varianta D^4_t má pozvolnější nástup způsobený různými efektivními poloměry, jejichž vliv je v D^4_{circ} kompenzován. Nižší hodnota příznaku je pak dána téměř dvojnásobnou střední hodnotou rozdělení D^4_t oproti D^4_{circ}.

[7]Kolmogorovův-Smirnovův test je popsán v sekci 4.1.1.



Obrázek 3.9: Porovnání dvou variant příznaku D^4 - příznaků D^4_{circ} a D^4_{diskr} pro obrázek 1.3



Obrázek 3.10: Porovnání dvou variant příznaku D^4 - příznaků D^4_{circ} a D^4_t pro obrázek 1.3

Z uvedených výsledků pro příznak D^4 nelze jednoznačně říci, jestli je některá z jeho variant je lépe či hůře svázaná s velikostí objektů. Z porovnání variant D^4_{circ} a D^4_{diskr} pak vidíme, že pro objekty blízké kruhovým můžeme místo výpočetně náročné varianty D^4_{diskr} používat kruhovou aproximaci D^4_{circ}, aniž bychom zanesli do výsledků výrazné nepřesnosti. Tyto závěry bohužel nelze aplikovat na pozdější fáze růstu tenkých vrstev. Pokud by byly objekty příliš nepravidelné, mohly by se kruhy, jimiž aproximujeme tvar objektů v D^4_{circ}, překrývat.

Zdrojové kódy, které vytvářejí obrázky (nejen) v této kapitole či počítají hodnoty příznaků nejsou obsaženy v této práci. Stejně tak zdrojové kódy naprogramovaných modelů. Všechny jsou ale nahrány na kompaktním disku přiloženému k této diplomové práci. Zdrojové kódy, stejně jako celá diplomová práce pak bude přístupná na autorově stránce `http://jindra.matfyz.cz`. Algoritmy byly naprogramovány v jazyce R verze 2.12.1 (2010-12-16) [14]. Část kódů byla rovněž naprogramována v programovacím jazyku Java. Ke generování Voronoiova dláždění a Delaunayovy triangulace byl využita knihovna EBImage určená pro jazyk R, která je volně dostupná například na `http://www.bioconductor.org/packages/2.9/bioc/src/contrib/EBImage_3.9.8.tar.gz`. V Javě jsme generoval Voronoiovo dláždění a Delaunayovu triangulaci pomocí balíku jdt dostupného volně na `http://code.google.com/p/jdt/`.

Kapitola 4

Statistické metody

V počítačové fyzice potřebujeme často zpracovávat různé soubory dat získané pomocí simulací. Snahou je z nasbíraných dat určit co nejpřesněji klíčové parametry či závislosti. Přesnost může být ovlivněna množstvím nasbíraných dat, ale i způsobem zpracování. U některých statistických metod je pak možné odhadnout i nepřesnost hledané veličiny.

Tato kapitola obsahuje souhrn statistických metod, které jsou v této práci ke zpracování dat použity. První část kapitoly se věnuje statistickým testům. Následuje přehled metod sloužících k odhadům hustoty pravděpodobnosti. Kapitolu pak uzavírá sekce popisující metodu Bootstrap.

4.1 Statistické testy

Statistické testy se používají k otestování platnosti nějaké (předem určené) hypotézy. Hypotéza je tvrzení o veličině, jejíž pravděpodobnostní rozdělení neznáme. Příkladem hypotézy je tvrzení „střední hodnota dané veličiny je rovna m“. Pravdivost této hypotézy testujeme na základě souboru hodnot dané veličiny.

Statistický test je pravidlo, které nám na základě našeho souboru dat přiřadí jedno z rozhodnutí zamítnout či nezamítnout původní (tzv. nulovou) hypotézu H_0. Zamítnutí či nezamítnutí pak v našem případě interpretujeme následujícím způsobem. Říkáme, že rozdíl střední hodnoty souboru od hodnoty m je/není větší, než by bylo možné očekávat v důsledku náhodné variability. Nulovou hypotézu tedy zamítáme, když je uspořádání dat v našem souboru velmi nepravděpodobné za předpokladu, že platí nulová hypotéza. Tuto (ne)pravděpodobnost nám měří tzv. p-hodnota (p-value), která je definovaná jako pravděpodobnost, že při platnosti nulové hypotézy byl náhodně vygenerován právě tento soubor dat ($p = P(data|H_0)$).

Konvencí je zamítat hypotézu, pokud je p-hodnota menší než $\alpha = 0{,}05$. Hodnotu α nazýváme hladinou významnosti.

V této práci budeme využívat dva statistické testy - Kolmogorovův-Smirnovův test shodnosti rozdělení a Shapirův-Wildův test normality. Spolu s těmito dvěma testy je v této sekci popsán i tzv. Q-Q diagram.

4.1.1 Kolmogorovův-Smirnovův test

Kolmogorovův-Smirnovův test (dále jen KS test) nám umožňuje testovat, zda dva soubory náhodných veličin pocházejí ze stejného pravděpodobnostního rozdělení. KS test nám také umožňuje testovat, zda jedna sada dat je realizací určitého jevu se známým (předem určeným) pravděpodobnostním rozdělením. V obou případech musí být data jednorozměrná.

Test je založen na porovnání empirických distribučních funkcí v prvním případě a na porovnání distribuční funkce a empirické distribuční funkce v případě druhém. Posuzovanou veličinou je maximální rozdíl. Tato veličina se nazývá Kolmogorovova-Smirnovova statistika a je definovaná jako

$$D_{m,n} = \sup_x |F_{1,m}(x) - F_{2,n}(x)| \,,$$

resp.

$$D_n = \sup_x |F_n(x) - F(x)| \,.$$

Nulovou hypotézu zamítáme, pokud

$$\sqrt{\frac{mn}{m+n}} D_{m,n} > K_\alpha \,,$$

resp. pokud

$$\sqrt{n} D_n > K_\alpha \,,$$

kde $K(x)$ je Kolmogorova distribuce a K_α je α-tý kvantil této distribuce. Kvantily Kolmogorovovy distribuce nejsou analyticky vyjádřitelné, je nutné počítat je numericky. Pro dostatečně velké soubory dat lze použít asymptotický tvar distribuce. Pro dostatečně velké m a n (větší než 100) můžeme použít pro $\alpha = 0{,}05$ hodnotu $K_\alpha \doteq 1{,}3581$. V práci je použita varianta KS testu obsažená v programu R [14], která počítá hodnoty Kolmogorovovy distribuce podle [15] pro malé soubory dat a podle článku [16] pro velké, jak je uvedeno v programové dokumentaci.

4.1.2 Shapirův-Wilkův test

Shapirův-Wilkův test je určen k testování normality. Test nám říká, jak je pravděpodobné, že daná sada dat pochází z normálního rozdělení. Klíčovou veličinou je W-statistika definovaná jako

$$W = \frac{(\sum_{i=1}^n a_i x_{(i)})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \,,$$

kde $x_{(i)}$ je i-tá nejmenší hodnota v souboru dat a a_i jsou koeficienty, které jsou odvozeny ze středních hodnot a varianční matice pořádkových statistik náhodného výběru z N(0,1) o rozsahu n. Tyto koeficienty bývají počítány numericky.

Pro praktické výpočty byla použita verze testu obsažená v programu R implementující algoritmus uvedený v článku [17]. Tento test budeme využívat v sekci 4.2.4.

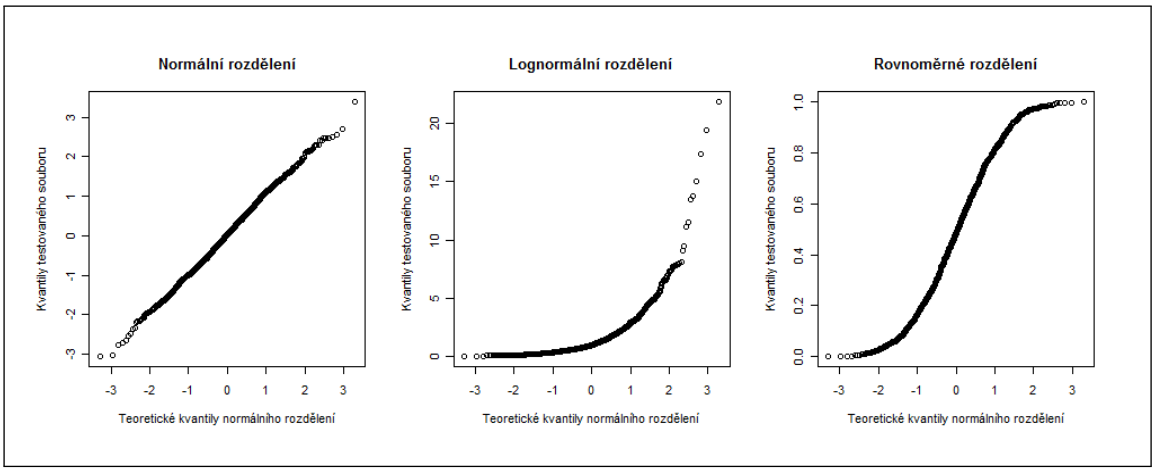
4.1.3 Q-Q diagram

Kvantilově-kvantilový diagram (dále jen Q-Q diagram) není statistickým testem, ale formou zobrazení dat. Můžeme pomocí něho posoudit, nakolik jsou si dvě distribuce podobné. Bod v grafu odpovídá jednomu kvantilu[1] druhé distribuce (y-ová souřadnice) vynesenému proti stejnému kvantilu první distribuce (x-ová souřadnice). V grafu je tedy vynesena křivka parametrizovaná hodnotou kvantilu. Pokud budou dvě porovnávané distribuce blízké, bude se výsledná křivka podobat funkci $y = x$.

Q-Q diagram lze využít i k posouzení, zda náhodný výběr byl vygenerovaný z určitého rozdělení. V konečném souboru dat bude jen n různých hodnot kvantilů, do grafu tedy vyneseme pouze ty. Diagram pak bude místo parametrické křivky obsahovat jen n bodů. Pokud budou ležet body blízko funkce $y = x$, je pravděpodobné, že jsou vygenerovány z tohoto rozdělení. Pokud budou vynesené hodnoty daleko od této křivky, je vysoce pravděpodobné, že nepocházejí z porovnávané distribuce.

Pomocí Q-Q diagramu můžeme posuzovat i to, jestli soubor dat pochází z normálního rozdělení bez ohledu na jeho střední hodnotu a rozptyl. Stačí vynést data oproti normálnímu rozdělení N(0,1) a následně provést lineární transformaci os tak, aby vynesené hodnoty byly co nejblíže funkci $y = x$. Z transformačních koeficientů můžeme dopočíst nejpravděpodobnější střední hodnotu a rozptyl.

Na obrázku 4.1 vidíme ilustraci Q-Q diagramu. V obou případech testujeme, jestli je soubor dat vygenerovaný z normálního rozdělení. V prvním případě tomu tak je, v druhém jsme použili lognormální rozdělení a třetí soubor byl vygenerován z rovnoměrného rozdělení na intervalu $[0, 1]$.



Obrázek 4.1: Ilustrace Q-Q diagramu

Q-Q diagram je velice užitečný nástroj na první zhodnocení, neměl by však nahrazovat statistické testování.

[1] Tzv. $(\alpha \cdot 100)$ní kvantil x_α je taková hodnota, která dělí všechny hodnoty náhodné veličiny či souboru dat na dvě části tak, že přibližně $(\alpha \cdot 100)\%$ naměřených hodnot je menších nebo rovných tomuto kvantilu a zbylých přibližně $((1 - \alpha) \cdot 100)\%$ naměřených hodnot je větších nebo rovných tomuto kvantilu.

4.2 Odhady hustoty pravděpodobnosti

V předchozí kapitole jsme se v několika případech dostali do situace, kdy jsme se na základně souboru dat pokoušeli odhadnout hustotu pravděpodobnosti rozdělení, ze kterého byla data vygenerována. V této části uvedeme přehled metod, které tuto úlohu řeší – na základě dostupných dat $\{x_i\}_{i=1}^N$ odhadují hustotu pravděpodobnosti ρ. Zmíníme výhody a nevýhody jednotlivých metod a ukážeme si jejich aplikaci na vzorových úlohách. Ukážeme si na tom, že výběr vhodné metody může zásadně ovlivnit přesnost odhadu.

4.2.1 Rozdělení metod

Pokud nemáme žádné apriorní informace o hustotě pravděpodobnosti, používáme při jejím odhadování pouze soubor dat $\{x_i\}_{i=1}^N$. Odhad pak bude záviset jen na datech, která máme k dispozici ($\hat{\rho}(x) = f(x, \{x_i\}_{i=1}^N)$. Takovýto odhad nazýváme odhadem neparametrickým. Příkladem takovéto metody je histogram, frekvenční polynom či metoda jádrových odhadů.

V některých případech můžeme předpokládat, že rozdělení patří do některé třídy pravděpodobnostních rozdělení (například že hledané rozdělení je normální). Pak se snažíme co nejpřesněji odhadnout parametry, jimiž je už pravděpodobnostní rozdělení určeno přesně (u normálního rozdělení střední hodnotu a rozptyl). Odhad hustoty pravděpodobnosti tedy závisí jen na odhadech parametrů $\theta = (\theta_1, \theta_2, \dots)$, tedy dostáváme $\hat{\rho}(x) = f(x, \hat{\theta}(\{x_i\}_{i=1}^N))$. V tomto případě jde o parametrické odhady jejichž příkladem je metoda maximální věrohodnosti.

Neparametrické metody jsou do velké míry univerzální. Nepotřebují žádné silné apriorní předpoklady o hledaném pravděpodobnostním rozdělení. Parametrické metody se dají použít pouze v některých případech. V případech, kdy je jejich použití oprávněné (hustota pravděpodobnosti opravdu patří do dané třídy pravděpodobnostních rozdělení), dávají mnohem přesnější výsledky než odhady neparametrické. Pokud však předpoklady neplatí, není takový parametrický odhad konzistentní (není zajištěna konvergence odhadu k přesnému řešení se zvyšujícím se počtem dat).

4.2.2 Kritéria přesnosti odhadu

Než přejdeme k jednotlivým metodám, musíme si zvolit kritéria, podle kterých budeme posuzovat, jak dobrý odhad daná metoda dává. Jednou z možností je porovnávat jeden konkrétní odhad $\hat{f}$ se skutečnou f. To se nám bude hodit v numerických experimentech. Druhou možností je zaměřit se na střední hodnotu odhadu $E(\hat{f})$ a porovnávat ji se skutečnou hustotou pravděpodobnosti.

V prvním případě nám nejlépe poslouží L^p-normy rozdílu $\hat{f}$ a f, speciálně pak L^2-norma, která je vhodná k teoretickým úvahám. Ve shodě s literaturou budeme tuto normu značit ISE - Integrated square error:

$$ISE = \int (\hat{f} - f)^2 \mathrm{d}x \,,$$

kde integrujeme přes celou reálnou osu[2].

V druhém případě je jednou z možností porovnat odhad a skutečnou funkci v každém bodě. Nejčastěji se porovnává druhá mocnina rozdílu hodnot

$$MSE(x) = E[\hat{f}(x) - f(x)]^2 = \mathrm{Var}\{\hat{f}(x)\} + \mathrm{Bias}^2\{\hat{f}(x)\}\,.$$

Zkratka MSE znamená Mean Square Error – střední kvadratická chyba, přičemž středování probíhá přes různé realizace odhadu. Člen $\mathrm{Var}\{\hat{f}(x)\}$ udává rozptyl pro odhad, zatímco člen $\mathrm{Bias}^2\{\hat{f}(x)\} = E[\hat{f}] - f(x)$ nám dává systematickou chybu odhadu.

Zintegrováním kritéria MSE dostáváme kritérium IMSE (Integrated Mean Square Error)

$$IMSE = \int MSE(x)\mathrm{d}x = \int E[\hat{f}(x) - f(x)]^2\mathrm{d}x =$$
$$= E\int [\hat{f}(x) - f(x)]^2\mathrm{d}x = E[ISE] =: MISE,$$

kde střední hodnotu a integrál jsme mohli zaměnit díky Fubinniho větě. Kritérium $MISE$ nám tedy určuje střední hodnotu L^2-normy rozdílu odhadu a funkce. Zároveň nám dává integrál ze střední kvadratické chyby.

Na závěr se podívejme na další přirozenou vlastnost, kterou bychom od dobré metody očekávali. Jde o konvergenci chyby k nule při zvětšování souboru dat. Tuto vlastnost popisuje pojem konzistence.

Řekneme, že metoda je konzistentní kvadraticky v průměru (consistent in the mean square), pokud $MSE(x) \to 0$ pro $N \to \infty$.

Dále nás bude zajímat rychlost konvergence, kterou budeme vyjadřovat Landauovým symbolem $\mathcal{O}$. Řekneme, že $f_n = \mathcal{O}(g_n)$, pokud $\exists c > 0 : f_n/g_n \to c$ pro $n \to \infty$.

4.2.3 Neparametrické odhady

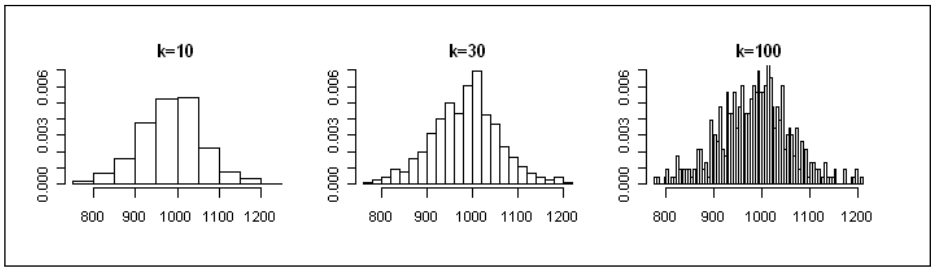
Histogram

Mějme soubor nezávislých náhodných jevů $\{x_i\}_{i=1}^N$. Nejjednodušší způsob, jak odhadnout pravděpodobnostní rozdělení tohoto jevu, je vynést data do tzv. histogramu. Ten pak bude naším odhadem hustoty pravděpodobnosti.

Histogram sestrojíme takto: Vezmeme interval I, ve kterém se vyskytují všechny hodnoty x_i, a rozdělíme ho (obvykle ekvidistantně) na k podintervalů – tzv. binů $\{I_i\}_{i=i}^k$, $I_j \cap I_k = \emptyset$ pro $\forall j \neq k$, $\bigcup_{i=1}^k I_i = I$. Spočítáme četnosti n_i (počty hodnot, které se vyskytují v i-tém binu). Histogramem pak budeme rozumět po částech konstantní funkci, která bude nulová mimo interval I a na každém podintervalu I_i bude mít hodnotu $\frac{n_i}{N}$ (relativní četnost).

Je zřejmé, že histogram závisí nejen na hodnotách $\{x_i\}_{i=1}^N$, ale též na volbě intervalu I a na volbě dělení $\{I_i\}_{i=1}^k$. Pokud se omezíme na ekvidistantní dělení a za okraje intervalu vezmeme krajní hodnoty souboru, stále nám ještě zbývá určit parametr k (počet binů). Podívejme se, jak může počet podintervalů ovlivnit vzhled histogramu.

[2]Zde nastává problém ve značení. Některá rozdělení mají neomezený nosič, zatímco jiná jsou definovaná pouze na určitém intervalu. Abychom zachovali obecnost vzorců, budeme u integrálů vynechávat meze.



Obrázek 4.2: Ilustrace vlivu počtu binů na vzhled histogramu

Na obrázku 4.2 vidíme tři různé histogramy pro stejná data[3]. Histogramy se liší jen počtem binů. Vidíme, že v prvním případě histogram zakrývá veškeré detaily, které jsou vidět v druhých dvou. U třetího histogramu naopak vidíme celkem vysokou fluktuaci hodnot, danou malým počtem hodnot v binech. Opticky vypadá ze tří zde uvedených histogramů nejlépe ten prostřední. Je to však skutečně tak? Lze volit počet binů automaticky podle nějakého pravidla a existuje vůbec ideální počet binů?

Nyní uvedeme stručný přehled nejvíce používaných pravidel pro volbu počtu binů, který nám tyto otázky alespoň částečně zodpoví. Tento přehled nebude kompletní ani aktuální, protože problematika je rozsáhlá a v současnosti se stále vyvíjejí nové metody. Výklad bude chronologicky postupovat od prvního pravidla odvozeného roku 1926 Sturgesem až po metody z počátku devadesátých let.

Patrně nejvíce používaný je vztah

$$k_{\mathrm{St}} = 1 + \log_2 N$$

odvozený za předpokladu normality rozdělení již zmiňovaným Sturgesem [19]. Tento odhad je díky své jednoduchosti používán velice často, přestože je známo [20], že dává dobré výsledky jen pro malé soubory dat ($N \approx 100$). Pro obsáhlejší soubory dává příliš malé počty binů a histogram tak zakrývá detaily.

Šířku intervalu I a tedy i šířku binu (kterou budeme dále značit jako h) ovlivňují krajní hodnoty souboru. Problém nastává hlavně u rozdělení, jejichž hustota pravděpodobnosti má neomezený nosič. Proto se častěji pracuje s pravidly určujícími šířku binu namísto počet binů na daném intervalu. Sturgesův odhad pro k lze (pokud uvažujeme normální rozdělení) jednoduše převést [20] na odhad pro šířku binu

$$h_{\mathrm{St}} = \frac{2\sigma\sqrt{2\log_2(N/2)}}{1 + \log_2 N}. \tag{4.1}$$

Sturgesovo pravidlo nám nedává žádné záruky pro minimalizaci chyby. Další odhady už vesměs předpokládají, že hledaná hustota pravděpodobnosti je na každém binu lipschitzovská. (Funkce f je lipschitzovská na intervalu I_i, pokud $\exists C > 0 : |f(u) - f(v)| \leq C|u - v| \quad \forall u, v \in I_i$.) Analýzou MSE pak lze dokázat [21], že histogram je konzistentním (kvadraticky v průměru) odhadem hustoty pravděpodobnosti,

[3]Byla použita data z článku [18]. Pomocí histogramu jsou vyneseny informace o stáří lidí (v měsících) přicházejících do domova důchodců v Pale Alto v Kalifornii.

pokud volba h zajistí, aby $N \to \infty$ implikovalo $h \to 0$ a $Nh \to \infty$. Sturgesovo pravidlo je tedy konzistentní odhad.

V roce 1979 ukázal Scott [22], že rozepsáním $MISE$ a zanedbáním vyšších řádů Taylorova rozvoje ρ můžeme dostat asymptoticky optimální volbu šířky binu takto

$$h^*_{\mathrm{Sc}} = \left(\frac{6}{NR(\rho')} \right)^{1/3}, \tag{4.2}$$

kde $R(\rho') = \int (\rho'(x))^2 \mathrm{d}x$.

Toto pravidlo minimalizuje $AMISE$ ($MISE$ se zanedbáním vyšších řádů). Řád konvergence chyby je $\mathcal{O}(N^{-2/3})$.

Člen $R(\rho')$ však neznáme. Následující pravidla se snaží ho co nejlépe odhadnout. Prvním takovým odhadem je

$$h^{\mathrm{Gauss}}_{\mathrm{Sc}} = 3{,}5\sigma N^{-1/3}, \tag{4.3}$$

který používá pro odhad členu $R(\rho')$ normální rozdělení, pro něž $R(\rho'_{Gauss}) = 3,5\sigma$. Znak σ značí směrodatnou odchylku souboru dat.

Podobné pravidlo navrhli roku 1981 Freedman a Diaconis [23]:

$$h_{\mathrm{FD}} = 2\,\mathrm{IQ}\,N^{-1/3}\,. \tag{4.4}$$

Znak IQ (inter-quartile range) značí vzdálenost mezi prvním a třetím kvartilem (25. a 75. kvantilem).

Shrňme si, co nám předchozí výsledky říkají. První skutečností je, že šířka binu má být úměrná $N^{-1/3}$. Při takovéto volbě pak L^2-norma chyby klesá s rostoucím počtem dat přibližně jako $N^{-2/3}$. Pro ilustraci, zdvojnásobením velikosti souboru nám klesne L^2-norma chyby přibližně o 37 procent.

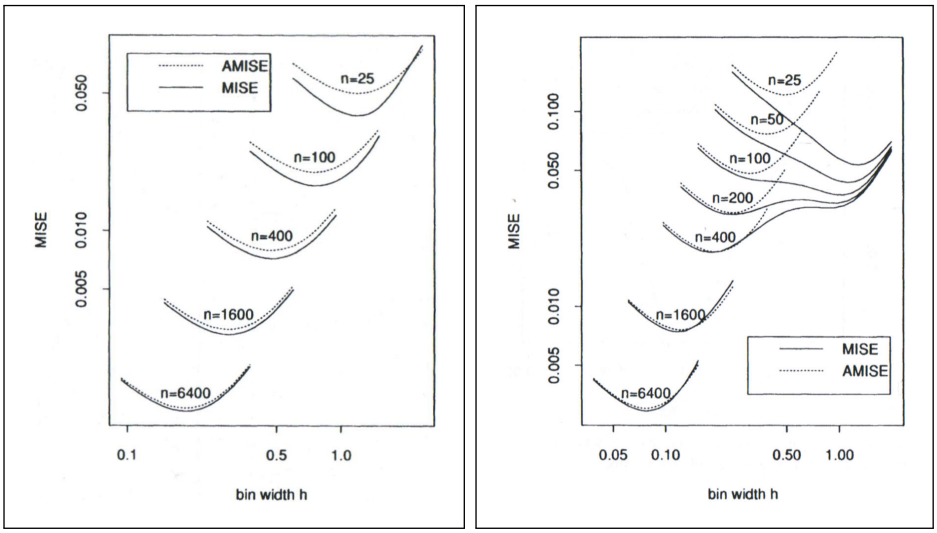
Optimální h můžeme spočítat, pokud budeme přesně znát samotnou odhadovanou funkci. Většinou ho můžeme pouze lépe či hůře odhadnout a od optimálního h^* se více či méně odchýlíme. Následující vztah [21] nám říká, jak se mění $AMISE$, pokud místo ideální šířky binu vezmeme jinou šířku vyjádřenou jako jeho násobek ch^*

$$\frac{AMISE(ch^*)}{AMISE(h^*)} = \frac{2 + c^3}{3c}\,.$$

Histogram tedy není příliš citlivý na malé výchylky šířky binu okolo optima.

Poslední poznámkou v tomto shrnutí je porovnání $MISE$ a $AMISE$ pro malé soubory dat. Zatímco pro $N \to \infty$ jde $(MISE - AMISE) \to 0$, u malých souborů se nemusí shodovat. Na obrázku číslo 4.3 vidíme porovnání pro normální a lognormální rozdělení pro soubory dat o velikostech $N = 25, 100, 400, 1\,600$ a $6\,400$. Vidíme, že zatímco u normálního rozdělení nám asymptotické odhady dávají dobrou shodu i pro malý soubor dat, u lognormálního dosahujeme dostatečné shody až u $N > 500$. Na předchozí odhady se tedy nemůžeme spolehnout u malých souborů dat a musíme mít vždy na paměti jejich asymptotický charakter.

Vraťme se k odhadům pro šířku binu. Roku 1985 pro ni dokázali Terrel a Scott [24] najít horní odhad. Pro pravděpodobnostní rozdělení s nosičem na $[-0{,}5; 0{,}5]$ je



Obrázek 4.3: Porovnání MISE a AMISE pro normální (vlevo) a lognormální (vpravo) rozdělení [21]

funkcionál $R(\rho')$ minimální pro funkci $\rho(x) := \rho_1(x) = \frac{3}{2}(1-4x^2)I_{[-0,5;0,5]}(x)$, kde $I_{[a,b]}$ je charakteristická funkce intervalu $[a,b]$. Po dosazení do vztahu (4.2) dostáváme

$$h^*_{\mathrm{Sc}} = \left(\frac{6}{NR(\rho')}\right)^{1/3} \leq \left(\frac{4}{N}\right)^{1/3} =: h_{\mathrm{TS}}\,. \tag{4.5}$$

Pro rozdělení s neomezeným nosičem, která mají rozptyl σ^2, minimalizuje člen $R(\rho')$ funkce $\rho_2(x) = \frac{15}{16\sqrt{7}\sigma}(1-\frac{x^2}{7\sigma^2})^2 I_{[-\sqrt{7}\sigma,\sqrt{7}\sigma]}(x)$ [25]. Po dosazení do vztahu (4.2) dostáváme

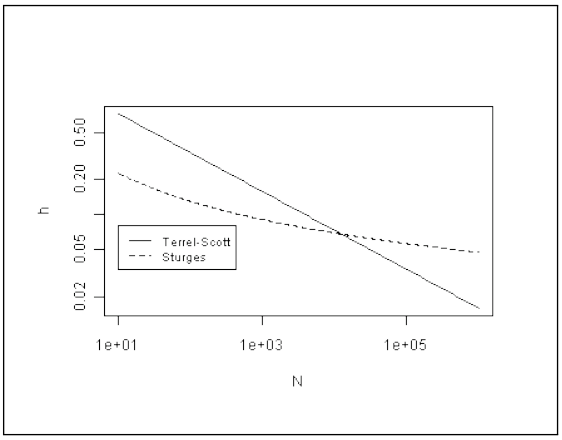
$$h^*_{\mathrm{Sc}} = \left(\frac{6}{NR(\rho')}\right)^{1/3} \leq \left(\frac{686\sigma^3}{5\sqrt{7}N}\right)^{1/3} \approx 3{,}729\sigma N^{-1/3} =: h_{\mathrm{T}}\,.$$

Zatímco horní odhad šířky binu známe, obecný dolní odhad znám není (člen $R(\rho')$ může růst nade všechny meze).

Nyní vidíme, proč Sturgesovo pravidlo není dobrou volbou pro velké soubory dat. Na obrázku 4.4 máme vynesené hodnoty h_{St} a h_{TS} pro různé velké soubory dat. Je vidět, že Sturgesovo pravidlo dává pro velké N větší šířku binu, než je horní odhad Terrela a Scotta. Histogram tedy bude nutně přehlazený.

V praxi se tedy můžeme buď spokojit s odhadem (4.5), nebo příslušný člen nějakým způsobem odhadnout s ohledem na informace, které máme k dispozici. O to se pokoušejí metody založené na tzv. krosvalidaci (Cross-Validation), kde se odhad derivace (a tedy i členu $R(\rho')$) získává odhadem z histogramu samého. Derivaci zde aproximují diference. Pokud si označíme hodnotu histogramu na intervalu i jako υ_i, pak $R(\rho')$ můžeme odhadnou jako

$$\hat{R}_{CV}(\rho') = \frac{1}{N^2h^3}\sum_k(\upsilon_{k+1}-\upsilon_k)^2 - \frac{2}{Nh^3}\,.$$



Obrázek 4.4: Porovnání Sturgesova pravidla a horního odhadu Terrela a Scotta

Dosazením do odhadu AMISE dostáváme

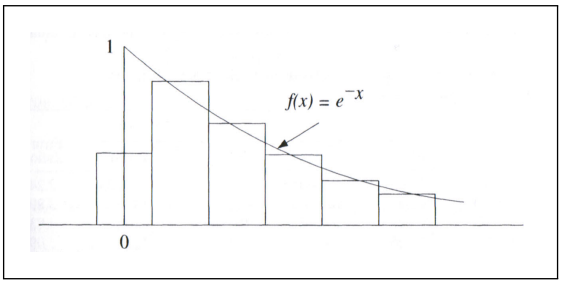
$$AMISE_{CV}(h) = \frac{5}{6Nh} + \frac{1}{12N^2h}\sum_k (v_{k+1} - v_k)^2.$$

Jako optimální šířka binu se pak bere takové h, které minimalizuje $AMISE_{CV}$.

Tento přístup má několik problémů. Prvním z nich je vysoká výpočetní náročnost. Druhým problémem je, že takto získaná $AMISE_{CV}(h)$ má často několik minim, a to včetně limitních případů $h = 0, h = \infty$.

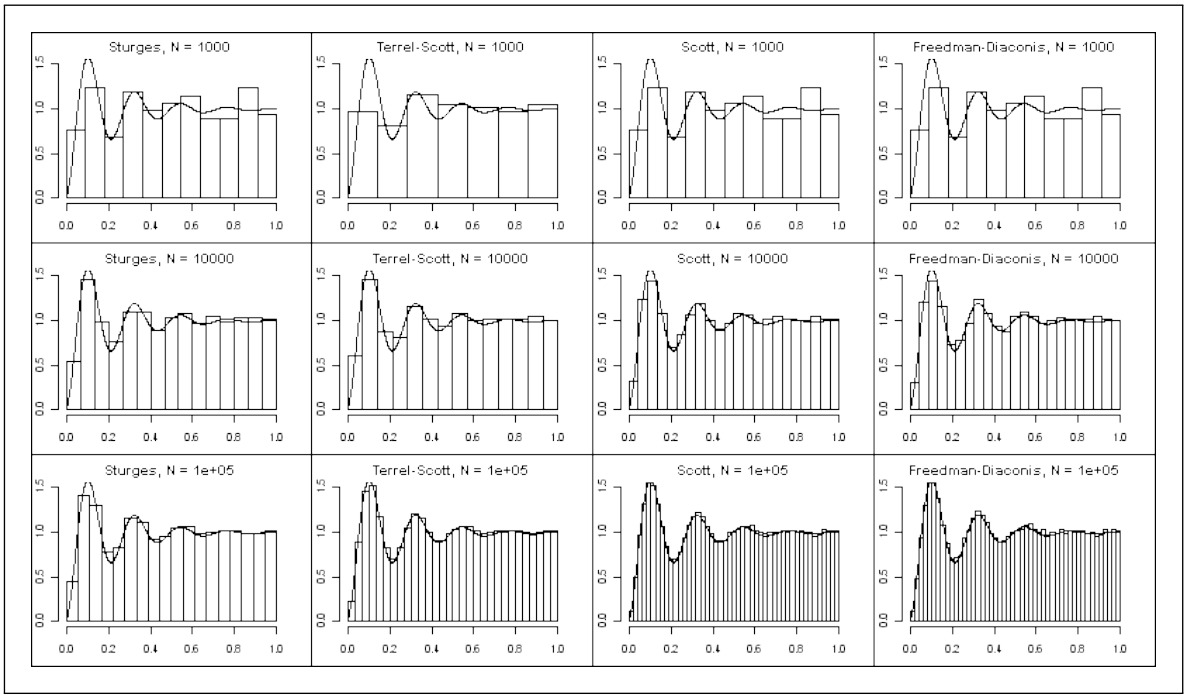
Dalším používaným přístupem je použít pro tvorbu histogramu neekvidistantní dělení. V místech, kde je více hodnot, by se histogram zjemnil a naopak v místech s menší počtem naměřených hodnost by se biny roztáhly. Tímto případem se dále nebudeme zabývat.

Poslední poznámkou je citlivost histogramu na posunutí binů. Při odvozování pravidel pro šířku binu jsme předpokládali lipschitzovskost hustoty pravděpodobnosti na každém binu. Pokud by však skutečná hustota pravděpodobnosti byla nespojitá a měla v jednom místě skok, podmínka lipschitzovskosti by byla splněna jen pokud by v místě skoku přecházel jeden bin v druhý. V opačném případě se nám zvýší chyba aproximace. Situaci dobře ilustruje obrázek 4.5.



Obrázek 4.5: Ilustrace vlivu nespojitosti rozdělení na chybu aproximace [21]

Z uvedeného souhrnu je vidět, že neexistuje jeden univerzální správný přístup,



Obrázek 4.6: Srovnání čtyř pravidel pro volbu šířky binu histogramů

jak volit šířku binu u histogramů. Všechna pravidla se snaží minimalizovat L^2-normu chyby. Pokud bychom chtěli minimalizovat nějakou jinou normu, dostali bychom odlišné odhady (např. při minimalizaci L^∞-normy chyby bude $h^* = \mathcal{O}(\log(n)n^{-1/3})$) [21].

Na závěr uvedeme srovnání jednotlivých pravidel pro šířku binu pro pravděpodobnostní rozdělení s hustotou pravděpodobnosti $\rho_1(x) = 0{,}994(1 - e^{-5x}\cos(\frac{90}{\pi}x))I_{[0,1]}(x)$. Histogramy vidíme na obrázku 4.6, v následující tabulce jsou vypsané L^2-normy chyb pro jednotlivé případy.

N	Sturges	Terrel, Scott	Scott	Friedman, Diaconis
1 000	0,228	0,246	0,244	0,244
10 000	0,130	0,138	0,097	0,097
100 000	0,103	0,066	0,047	0,048

Tabulka 4.1: L^2-normy chyb pro různá pravidla šířky binu histogramů

Vidíme, že zatímco pravidla navržená Scottem a Friedmanem s Diaconisem se od sebe téměř neliší, Sturgesovo pravidlo se od nich velmi odchyluje, a to jak opticky, tak chybou, která je pro $N = 100\,000$ už více než dvojnásobná.

Histogram je robustní metoda, kterou lze použít téměř v každé situaci. V této práci ji používáme k získání představy o tom, jaký charakter má hustota pravděpodobnosti. K volbě šířky binu budeme používat pravidlo (4.4) (Friedman-Diaconis). Pro přesnější odhady hustoty pravděpodobnosti na malém intervalu budeme používat parametrické metody, které jsou citlivější.

Jádrové odhady

Histogram je nejjednodušší odhad hustoty pravděpodobnosti. Dává nám konzistentní odhad ve smyslu MISE. Pokud zvolíme optimální šířku binu, zaručuje nám řád konvergence L^2-normy chyby $\mathcal{O}(N^{-2/3})$. Dává nám ale pouze po částech konstantní funkci, což nám v některých případech nemusí vyhovovat.

Další nevýhodou histogramu je, že jsme upřednostnili určité dělení před všemi ostatními. Dostali jsme tak lepší odhad pro hodnoty uprostřed binu a horší na jeho okrajích.

Metoda jádrových odhadů (anglicky kernel density estimation) odstraňuje obě uvedené nevýhody. Umožňuje nám dělat libovolně hladký odhad a obchází preferenci určitého dělení tak, že počítá odhad v každém bodě jako vážený průměr hodnot z určitého okolí. Zásadní otázkou zde je, jak široké okolí máme vzít a jaké váhy budeme používat.

Jak uvidíme, pro jádrové odhady lze vybudovat obdobnou teorii jako pro histogramy. Získáme vyšší řád konvergence, cenou za to bude vyšší citlivost na špatnou volbu h. Teorie jádrových odhadů je ještě rozsáhlejší, než teorie histogramů. Krátký přehled v této sekci čerpá z knihy [21].

Definujme váhovou funkci

$$w = \begin{cases} 1/2 & x \in [-1,1] \\ 0 & \text{jinde} \end{cases},$$

pak můžeme jako odhad hustoty pravděpodobnosti vzít

$$\hat{\rho}(x) = \frac{1}{N\,h} \sum_{i=1}^{N} w\left(\frac{x - x_i}{h}\right).$$

Podívejme se, jakým způsobem daný vzorec funguje. Hodnotu v bodě x počítá tak, že sečte všechny hodnoty x_i, pro které $|\frac{x-x_i}{h}| \leq 1$, a pronásobí je vhodnou konstantou. Ve výsledku v každém bodě dostáváme stejnou hodnotu, jakou bychom dostali, kdyby byl středem binu histogramu.

Odhad můžeme ještě vylepšit, pokud zvýhodníme x_i blízká x a znevýhodníme vzdálenější. Místo váhové funkce w budeme brát vhodnou funkci K (Kernel – jádro). Pro jednoduchost se dále budeme zabývat jen nezápornými symetrickými jádry s jednotkovým prvním momentem ($K(x) \geq 0$, $\int_{-\infty}^{\infty} K(x)\mathrm{d}x = 1$, $K(-x) = K(x)$). Často používanými jádry jsou gaussovské ($K_{\mathrm{G}} = N(0,1)$ - normální rozdělení s nulovou střední hodnotu a jednotkovým rozptylem) a Epanechnikovo ($K_{\mathrm{Ep}}(x) = \frac{3}{4}(1-x^2)I_{[-1,1]}$). První dává nekonečně hladký odhad ρ, druhé minimalizuje $AMISE$ mezi všemi symetrickými nezápornými jádry.

Odhady pomocí jádrových odhadů můžeme formulovat též jako konvoluci

$$\left[\frac{\mathrm{d}\hat{F}_N}{\mathrm{d}x}\right] * K = \int_{\infty}^{\infty} \left[\frac{1}{n}\sum_{i=1}^{N} \delta(t - x_i)\right] K(x-t)\mathrm{d}t = \frac{1}{n}\sum_{i=1}^{N} K(x - x_i)\,,$$

kde $\hat{F}_N$ je empirická distribuční funkce (po částech konstantní funkce). Jádrové odhady tedy můžeme brát jako shlazování empirické hustoty pravděpodobnosti.

Stejně jako u histogramů lze najít asymptoticky optimální h, kterému zde říkáme šířka jádra. Odhady platí pro nezáporná symetrická jádra s omezeným nosičem a konečným rozptylem a pro hustoty pravděpodobnosti ρ, pro které je ρ'' absolutně spojitá a $\rho''' \in L^2$

$$\begin{aligned} AMISE(h) &= \frac{R(K)}{Nh} + \frac{1}{4}\sigma_K^4 h^4 R(\rho'')\,, \\ h^* &= \left(\frac{R(K)}{\sigma_K^4 R(\rho'')}\right)^{1/5} N^{-1/5}\,, \\ AMISE(h^*) &= \frac{5}{4}[\sigma_K R(K)]^{-4/5} R(\rho'')^{1/5} N^{-4/5}\,, \\ \frac{AMISE(ch^*)}{AMISE(h^*)} &= \frac{8+c^9}{9c}\,, \end{aligned} \tag{4.6}$$

kde $\sigma_K^2 = \int_{-\infty}^{\infty} x^2 K(x)\mathrm{d}x$.

Co z těchto vztahů vyplývá? První zjištění je, že chyba závisí na parametrech jádra. Za druhé dostáváme rychlejší konvergenci chyby k nule. Dále ze čtvrtého vztahu vidíme, že jádrové odhady jsou citlivější na špatnou volbu šířky jádra, než histogram na šířku binu. Posledním zjištěním je, že pro odhad optimálního h nyní potřebujeme znát druhou derivaci skutečné hustoty pravděpodobnosti, zatímco u histogramů jsme potřebovali první.

Rychlost konvergence lze dále zlepšit až k $\mathcal{O}(N^{-1})$ (řád konvergence pro neparametrické metody je principiálně zdola omezen Cramérovou-Raovou mezí, kterou nelze překročit). To zajistíme použitím jádra s více nulovými momenty (ztrácíme nezápornost). Současně se nám ale bude zvyšovat citlivost na správnou volbu a optimální h bude záviset na vyšších derivacích hustoty pravděpodobnosti. Jejich výhoda – řád konvergence – se projeví až na velmi obsáhlých souborech dat. Pro tuto práci tedy nemají význam.

Pro přesnost odhadu je u jádrových odhadů, stejně jako u histogramů, klíčová volba h. Nejjednodušší je použít jako referenci normální rozdělení (podobně jako ve vztahu (4.3)), pro které je člen $R(\rho'')$ roven $\frac{3}{8\sqrt{\pi}\sigma^5}$. Dostáváme odhad optimální šířky jádra

$$h^{*Gauss} = \left(\frac{8\sqrt{\pi}R(K)}{3\sigma_K^4}\right)^{1/5} \sigma N^{-1/5}\,, \tag{4.7}$$

kde za σ dosazujeme směrodatnou odchylku hledané hustoty pravděpodobnosti.

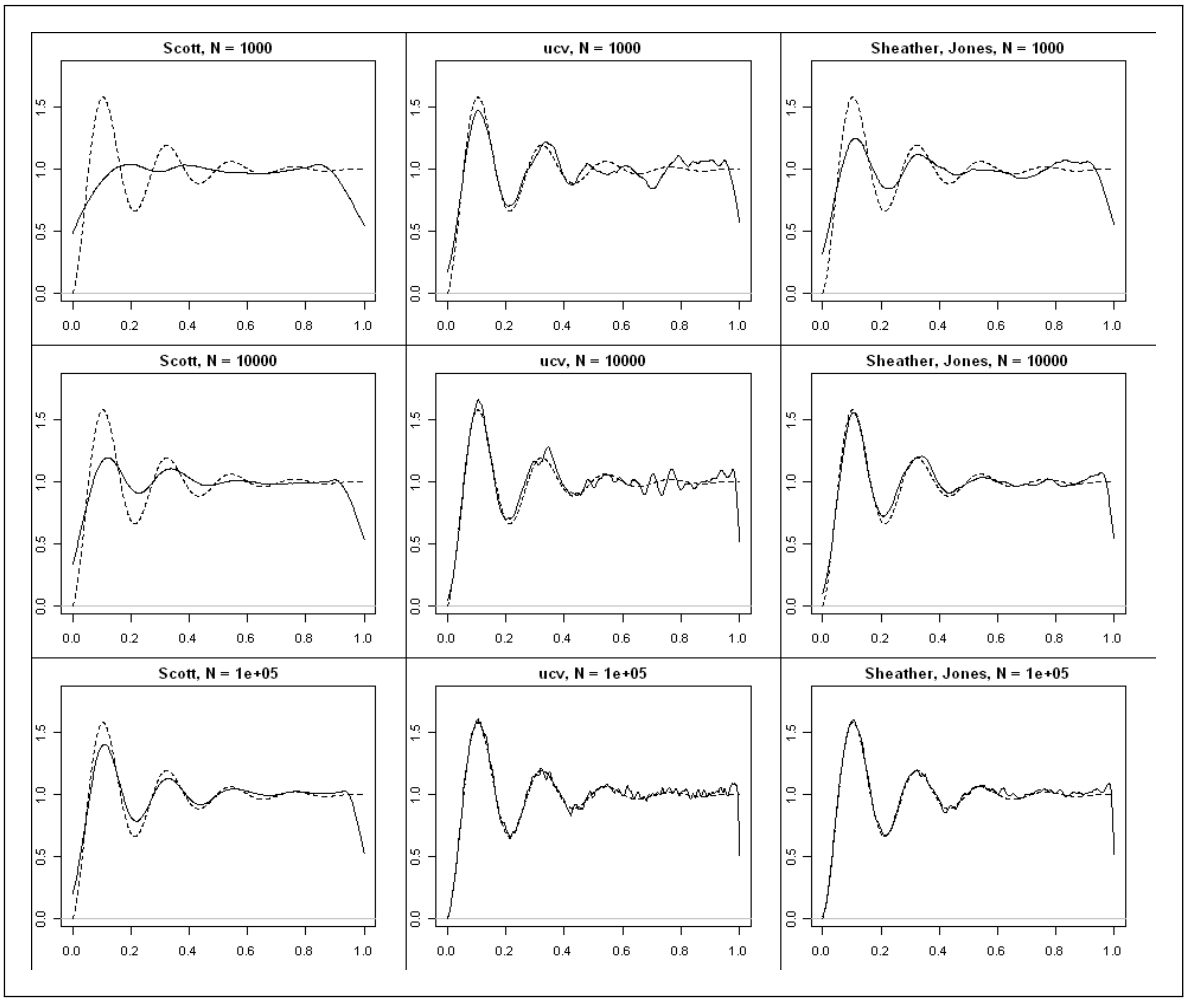
Opět můžeme najít horní odhad pro šířku jádra

$$h^{*max} = 3\left(\frac{R(K)}{35\sigma_K^4}\right)^{1/5} \sigma N^{-1/5}\,.$$

Podobně jako u histogramů i zde můžeme použít krosvalidaci k odhadu členu $R(\rho'')$. Odvození, postup a analýzu můžeme najít v článku [26], jeho vylepšenou verzi pak v článku [27].

Pro porovnání jsme zvolili pravidlo ze vztahu (4.7) (označujeme jako Scott), algoritmus unbiased cross-validation (značíme jako ucv) popisovaný v článku [26] a algoritmus

uvedený v článku [27], který označujeme jako Sheather, Jones podle autorů článku. Ve všech případech jsme používali Epanechnikovo jádro a konstantní šířku jádra na celém intervalu. Pravidla jsme opět testovali na náhodném výběru 1 000 (resp. 10 000 a 100 000) s pravděpodobnostním rozdělením $\rho_1(x) = 0{,}994(1 - e^{-5x}\cos(\frac{90}{\pi}x))I_{[0,1]}(x)$. Výsledky vidíme na obrázku 4.7.



Obrázek 4.7: Srovnání tří pravidel pro volbu šířky jádra (skutečná hustota pravděpodobnosti je vynesena čárkovaně)

První, čeho si musíme všimnout, je nepřesnost odhadů na začátku a na konci intervalu. Je způsobena nespojitostí hustoty pravděpodobnosti (resp. její derivace) na okrajích intervalu. Tato situace se dá uspokojivě vyřešit pomocí volby jiného tvaru jádra v okamžiku, kdy nosič jádra přesahuje okraje intervalu.

Další pozorování ukazuje, že ani jedno pravidlo nedokáže přesně aproximovat hustotu pravděpodobnosti na celém intervalu. Pravidlo Scott lépe aproximuje konec intervalu, ale na počátku je velice nepřesné a u malých souborů dat dokonce ani nevystihuje charakter hustoty pravděpodobnosti. Naopak pravidlo ucv dobře aproximuje hustotu pravděpodobnosti na začátku intervalu, ale na konci intervalu osciluje. Důvodem to-

hoto chování je vysoká citlivost jádrových odhadů na nepřesnou šířku jádra a měnící se charakter hustoty pravděpodobnosti. Pro první polovinu intervalu by bylo vhodnější užší jádro, pro druhou polovinu naopak širší.

V následující tabulce jsou uvedeny L^2-normy chyby pro jednotlivé případy.

N	Scott	ucv	Sheather, Jones
1 000	0,225	0,080	0,132
10 000	0,151	0,059	0,060
100 000	0,090	0,039	0,040

Tabulka 4.2: Srovnání L^2-normy chyby odhadu pomocí jádrových odhadů pro jednotlivá pravidla

Vzhledem k velké nepřesnosti odhadu na konci intervalu jsme spočítali i L^2-normy chyby rozdílu odhadu a skutečné hustoty pravděpodobnosti jen na intervalu $[0; 0{,}9]$, které jsou vyneseny v následující tabulce:

N	Scott	ucv	Sheather, Jones
1 000	0,219	0,068	0,121
10 000	0,140	0,047	0,038
100 000	0,071	0,027	0,024

Tabulka 4.3: Srovnání L^2-normy chyby odhadu na intervalu $[0; 0{,}9]$ pomocí jádrových odhadů pro jednotlivá pravidla

Pro ilustraci uvádíme ještě zvolené šířky jádra pro jednotlivé případy včetně optimálního h^* spočteného podle vzorce (4.6).

N	h_{Scott}	h_{ucv}	$h_{Sheather,Jones}$	h^*
1 000	0,0769	0,0251	0,0433	0,0489
10 000	0,0480	0,0097	0,0189	0,0308
100 000	0,0303	0,0034	0,0058	0,0195

Tabulka 4.4: Šířky jádra pro jednotlivá pravidla u jádrových odhadů

Vidíme, že pravidlo Scott šířku jádra nadsazuje, zatímco ucv dává zbytečně malou hodnotu. V dalších částech práce budeme pro jádrové odhady používat jen pravidlo Sheathera a Jonese v kombinaci s Epanechnikovým jádrem.

Pokud porovnáme přesnost histogramu a jádrových odhadů, vidíme, že přesněji aproximují hustotu pravděpodobnosti jádrové odhady. Chyba je řádově poloviční. Nevýhodou je však větší časová náročnost. Jádrové odhady budeme tedy používat pouze v případě, kdy nám půjde o co největší přesnost odhadu hustoty pravděpodobnosti.

4.2.4 Parametrické odhady

U neparametrických metod jsme nestanovovali žádné předpoklady pro hustotu pravděpodobnosti. Díky tomu jsme získali robustní metody použitelné v naprosté většině případů. Cenou za to byla malá přesnost odhadu a jeho pomalá konvergence k přesné hustotě pravděpodobnosti. V této části ukážeme, že pokud si vhodně omezíme množinu možných pravděpodobnostních metod, můžeme dostat přesnější odhad, který konverguje rychleji.

Nejprve uvedeme jednoduchý příklad. Představme si, že máme soubor hodnot vygenerovaných z normálního rozdělení s neznámou střední hodnotou a neznámým rozptylem. Pokud bychom neměli informaci o tom, že hledané rozdělení je normální, použili bychom neparametrické metody. S touto informací nám ale stačí co nejlépe určit pouze dva neznámé parametry, které hustotu pravděpodobnosti přesně definují.

Obecně budeme předpokládat, že rozdělení, které hledáme, patří do nějaké třídy pravděpodobnostních rozdělení $\mathcal{F} = \{f(x,\theta), f \text{ je hustota pravděpodobnosti pro všechna přípustná } \theta\}$. Předpokládáme tedy, že hustota pravděpodobnosti je určena jednoznačně až na konečný počet parametrů $\theta = (\theta_1, \theta_2, \ldots, \theta_M)$, $\theta \in \Psi \subset \mathbb{R}^M$. Cílem parametrických metod je najít optimální θ, které značíme θ^*. Odhad hustoty pravděpodobnosti tedy převádíme na hledání optimálního zástupce z dané třídy rozdělení – $\hat{\rho}(x) = f(x, \hat{\theta}(\{x_i\}_{i=1}^N))$.

Metoda maximální věrohodnosti

Metoda maximální věrohodnosti patří mezi parametrické odhady. Jejím cílem je vybrat ze zvolené třídy pravděpodobnostních rozdělení takové, které nejlépe odpovídá naměřeným datům $\{x_i\}_{i=1}^N$. Hledáme tedy takové θ, pro něž je nejpravděpodobnější, že by se při náhodném výběru z rozdělení $f(x,\theta)$ realizovala právě nám dostupná data $\{x_i\}_{i=1}^N$. Jelikož předpokládáme, že jednotlivé realizace jsou navzájem nezávislé a vybrané ze stejného rozdělení, hledáme $\hat{\theta} = \operatorname{argmax}_{\theta\in\Psi} \prod_{i=1}^N f(x_i,\theta)$.

Funkci $L(\theta, \{x_i\}_{i=1}^N) = \prod_{i=1}^N f(x_i,\theta)$ nazýváme **věrohodnostní funkcí** (L z anglického likelihood). Maximum této funkce hledáme za pomoci parciálních derivací:

$$\frac{\partial L}{\partial \theta_i} = 0 \quad \forall i = 1, 2, \ldots, M\,. \tag{4.8}$$

Této rovnici/sadě rovnic se říká **věrohodnostní rovnice**.

Hodnotu θ, kterou získáme jako řešení věrohodnostní rovnice (4.8), budeme označovat $\hat{\theta}^{ML}$ a budeme ji nazývat **maximálně věrohodným odhadem** parametru θ.

Často si můžeme výpočet zjednodušit tím, že místo věrohodnostní funkce maximalizujeme její přirozený logaritmus

$$l(\theta, \{x_i\}_{i=1}^N) := \log\left[L\left(\theta, \{x_i\}_{i=1}^N\right)\right] = \log\left[\prod_{i=1}^N f(x_i,\theta)\right] = \sum_{i=1}^N \log[f(x_i,\theta)]\,.$$

Díky tomu, že $f(x_i,\theta)$ jsou nezáporné a funkce logaritmus je monotónní funkcí, dostáváme stejný výsledek ($\operatorname{argmax}_{\theta\in\Psi} L = \operatorname{argmax}_{\theta\in\Psi} l$). Místo věrohodnostní funkce pak můžeme do věrohodnostní rovnice dosadit právě její logaritmus.

Ukažme si tuto metodu na již výše zmíněném příkladu s normálním rozdělením. Věrohodnostní funkce bude mít tvar $L(\mu, \sigma, \{x_i\}_{i=1}^N) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$, její logaritmus pak bude ve tvaru $l(\mu, \sigma, \{x_i\}_{i=1}^N) = \sum_{i=1}^N -\frac{(x_i-\mu)^2}{2\sigma^2} + \log(\frac{1}{\sqrt{2\pi\sigma^2}})$. V tomto případě dostáváme dvě věrohodnostní rovnice. První z nich $\sum_{i=1}^N \frac{2(x_i-\mu)}{2\sigma^2} = 0$ je splněna, pokud $\sum_{i=1}^N (x_i - \mu) = 0$, tedy pokud pro parametr μ použijeme odhad $\hat{\mu}^{ML} = \frac{1}{N}\sum_{i=1}^N x_i$. Odhad pro druhý parametr lze najít analogicky.

Hodnota odhadu vždy závisí na konkrétních hodnotách $\{x_i\}_{i=1}^N$, pro různé soubory dat (i kdyby byly vygenerovány z toho samého rozdělení) se věrohodnostní odhad může lišit. Odhad $\hat{\theta}^{ML}$ tak můžeme v závislosti na datech brát jako náhodnou veličinu s neznámým rozdělením. Z teorie víme (viz např. [28]), že při splnění vhodných předpokladů má tato veličina rozdělení asymptoticky normální se středem ve skutečné hodnotě. Navíc pro $N \to \infty$ jde rozptyl k nule, odhad je tedy konzistentní. Je také dokázáno, že mezi všemi asymptoticky nestrannými odhady jde rozptyl k nule asymptoticky nejrychleji (rychlost konvergence dosahuje Cramérovy-Raovy meze). Řečeno laicky, pokud budeme opakovat odhad několikrát, budou se odhadnuté hodnoty pohybovat kolem hodnoty skutečné, a čím větší navíc bude soubor dat, tím blíže se budou odhady nacházet.

Metodu maximální věrohodnosti zkusíme použít pro získání příznaku u radiální distribuční funkce. Budeme se snažit co nejlépe odhadnout pozici a výšku prvního maxima hustoty pravděpodobnosti. Nebudeme se pokoušet odhadnout celkový tvar hustoty pravděpodobnosti, ale jen příslušné části obsahující maximum. Na tomto úseku budeme předpokládat, že hustota pravděpodobnosti má tvar paraboly.

K otestování této úlohy použijeme rozdělení se známou hustotou pravděpodobnosti $\rho_1(x) = 0{,}994(1 - e^{-5x}\cos(\frac{90}{\pi}x))I_{[0,1]}(x)$. Jak už jsme uvedli, budeme chtít najít co nejpřesněji pozici x_{1max} a hodnotu y_{1max} prvního maxima hustoty pravděpodobnosti na základě souboru náhodných veličin vygenerovaných z tohoto rozdělení. Přibližnou pozici prvního maxima můžeme zjistit například pomocí histogramu či jádrového odhadu. Přibližná poloha maxima je okolo 0,1, proto si zvolíme pro aproximaci pomocí metody maximální věrohodnosti například interval $[x_l; x_r] = [0{,}06; 0{,}15]$. Za třídu funkcí $\mathcal{F}$ si zvolíme polynomy druhého stupně ($\mathcal{F} = \{f(x) = ax^2 + bx + c, x \in [x_l; x_r], a, b, c \in \mathbb{R}\}$). Dále budeme postupovat následovně:

Nejprve spočteme podíl hodnot spadajících do daného intervalu $\frac{N_{[x_l;x_r]}}{N}$. Tuto hodnotu použijeme jako odhad integrálu $\int_{x_l}^{x_r}(ax^2+bx+c)\mathrm{d}x$. Díky tomu můžeme eliminovat parametr c volbou

$$\hat{c} := \frac{-3bx_l - ^2 - 2ax_l^3 + 3bx_r^2 + 2ax_r^3 - 6\frac{N_{[x_l;x_r]}}{N}}{-6(x_r - x_l)}\,.$$

Následně budeme hledat hodnoty parametrů a, b takové, aby maximalizovaly výraz $\sum_{i:\, x_l \le x_i \le x_r} \log(ax_i^2 + bx_i + \hat{c})$. Řešení příslušných věrohodnostních rovnic nelze analyticky spočítat, proto budeme k nalezení optimálních a, b používat optimalizační metodu BFGS [29] (podle autorů Broyden-Fletcher-Goldfarb-Shanno). Jelikož pro některé hodnoty a, b původní výraz nemusí být definovaný, budeme namísto něho maximalizovat výraz

$$\begin{cases} \sum\limits_{i:\, x_l \le x_i \le x_r} \log(ax_i^2 + bx_i + \hat{c}) & \text{pokud } (ax_i^2 + bx_i + \hat{c}) > 0 \\ -\infty & \text{pokud } (ax_i^2 + bx_i + \hat{c}) \le 0\,. \end{cases}$$

Vyjádřeno slovně – pokud by se stalo, že pro nějaké x_i by byl výraz $ax_i^2 + bx_i + c$ záporný, položíme výraz rovný minus nekonečnu. Tím tyto případy penalizujeme. Za maximálně věrohodné odhady pozice a výšky prvního maxima budeme brát pozici vrcholu paraboly a hodnotu paraboly v tomto bodě: $\hat{x}_{1max}^{ML} = -\frac{b}{2a}$, $\hat{y}_{1max}^{ML} = -\frac{b^2}{4a} + c$.

Tuto metodu jsme testovali na náhodných výběrech o velikostech $N = 100$, 1 000, 10 000. Pro každou velikost výběru jsme postup opakovali K-krát, $K = 100$, 1 000, 10 000. Tím jsme dostali celkem devět souborů odhadů, které jsme dále statisticky zpracovali.

Nejprve jsme vyloučili outliery – hodnoty, které byly na první pohled špatně. K označení outlierů jsme používali kritérium založené na kvantilech souboru dat popsané v [30]. Pokud si označíme q_1 jako první kvartil[4] dat a q_3 jako třetí, pak za outlier budeme považovat hodnotu, která bude větší, než $q_3 + 1{,}5(q_3 - q_1)$ nebo menší než $q_1 - 1{,}5(q_3 - q_1)$. Outliery jsme hledali mezi hodnotami $\hat{x}_{1max}^{ML}$ a $\hat{y}_{1max}^{ML}$. V případě nalezení outlieru jsme vyloučil ze souboru všechny údaje týkající se tohoto odhadu (kromě odhadů samotných i hodnoty a, b, c a hodnotu věrohodnostní funkce L) a test jsem znovu aplikovali na zbylá data.

U takto očištěných dat jsme spočetli střední hodnotu a směrodatnou odchylku. Výsledky celé procedury jsou uvedeny v následující tabulce.

N\K	100	1 000	10 000
100	$\hat{x}_{1max}^{ML} = 0{,}1066 \pm 0{,}0159$ $\hat{y}_{1max}^{ML} = 1{,}6722 \pm 0{,}7623$ počet outlierů: 33	$\hat{x}_{1max}^{ML} = 0{,}1046 \pm 0{,}0189$ $\hat{y}_{1max}^{ML} = 1{,}7587 \pm 0{,}8923$ počet outlierů: 286	$\hat{x}_{1max}^{ML} = 0{,}1052 \pm 0{,}0178$ $\hat{y}_{1max}^{ML} = 1{,}7243 \pm 0{,}8566$ počet outlierů: 3093
1 000	$\hat{x}_{1max}^{ML} = 0{,}1057 \pm 0{,}0089$ $\hat{y}_{1max}^{ML} = 1{,}5504 \pm 0{,}1569$ počet outlierů: 28	$\hat{x}_{1max}^{ML} = 0{,}1052 \pm 0{,}0134$ $\hat{y}_{1max}^{ML} = 1{,}5626 \pm 0{,}1768$ počet outlierů: 278	$\hat{x}_{1max}^{ML} = 0{,}1050 \pm 0{,}0123$ $\hat{y}_{1max}^{ML} = 1{,}5731 \pm 0{,}1671$ počet outlierů: 2994
10 000	$\hat{x}_{1max}^{ML} = 0{,}1056 \pm 0{,}0036$ $\hat{y}_{1max}^{ML} = 1{,}5781 \pm 0{,}0563$ počet outlierů: 2	$\hat{x}_{1max}^{ML} = 0{,}1059 \pm 0{,}0033$ $\hat{y}_{1max}^{ML} = 1{,}5885 \pm 0{,}0545$ počet outlierů: 50	$\hat{x}_{1max}^{ML} = 0{,}1058 \pm 0{,}0034$ $\hat{y}_{1max}^{ML} = 1{,}5872 \pm 0{,}0552$ počet outlierů: 523

Tabulka 4.5: Souhrn odhadů polohy a výšky prvního maxima metodou maximální věrohodnosti

Skutečná hodnota x_{1max} je přibližně 0,1036, pro y_{1max} je 1,5867 (spočteno na základě analytického vyjádření hustoty pravděpodobnosti). Z údajů v tabulce vidíme, že skutečná hodnota ve všech případech bezpečně leží v intervalu daném střední hodnotou a směrodatnou odchylkou.

Nyní ověříme, že výsledné rozložení odhadů se blíží s rostoucím N k normálnímu rozdělení. Použijeme k tomu Shapirův-Wildův test normality. Pro daný soubor dat dává test hodnotu p (p-value), která určuje, jak je pravděpodobné, že by hodnoty

[4]Kvartily jsou hodnoty rozdělující soubor dat na čtvrtiny. První kvartil odpovídá 25. kvantilu, třetí kvartil 75. kvantilu.

souboru byly náhodně vygenerovány z normálního rozdělení. Hodnota p nám zároveň dává hladinu významnosti, na které můžeme vyloučit hypotézu tvrdící, že hodnoty souboru byly náhodně vygenerovány z normálního rozdělení. Hodnoty p značíme p_x^{SW} pro polohy a p_y^{SW} pro výšky.

N\K	100	1 000	10 000
100	$p_x^{SW} = 0{,}828$ $p_y^{SW} = 0{,}096$	$p_x^{SW} = 9{,}836 \cdot 10^{-3}$ $p_y^{SW} = 8{,}989 \cdot 10^{-7}$	$p_x^{SW} = 1{,}069 \cdot 10^{-9}$ $p_y^{SW} = 7{,}104 \cdot 10^{-18}$
1 000	$p_x^{SW} = 0{,}243$ $p_y^{SW} = 0{,}719$	$p_x^{SW} = 7{,}339 \cdot 10^{-6}$ $p_y^{SW} = 9{,}053 \cdot 10^{-2}$	$p_x^{SW} = 3{,}340 \cdot 10^{-20}$ $p_y^{SW} = 1{,}570 \cdot 10^{-6}$
10 000	$p_x^{SW} = 0{,}720$ $p_y^{SW} = 0{,}839$	$p_x^{SW} = 5{,}369 \cdot 10^{-2}$ $p_y^{SW} = 3{,}765 \cdot 10^{-2}$	$p_x^{SW} = 2{,}757 \cdot 10^{-7}$ $p_y^{SW} = 1{,}201 \cdot 10^{-5}$

Tabulka 4.6: Přehled p-hodnot Shapirova-Wildova testu pro soubory odhadů polohy a výšky prvního maxima metodou maximální věrohodnosti

Z tabulky vidíme, že ve většině případů hodnota p roste s rostoucím počtem bodů použitých k odhadu pomocí maximální věrohodnosti N, jak předpokládá teorie. Tento trend však není příliš výrazný. Zároveň vidíme, že test s rostoucím počtem odhadů bezpečněji poznává, že odhady nejsou rozloženy gaussovsky.

Nyní se podrobněji podívejme na případ $N = 100$, $K = 1000$. Na obrázcích 4.8 a 4.9 vidíme srovnání souborů odhadů pro pozici, resp. výšku prvního maxima. Na obou obrázcích jsou porovnány soubory původní se soubory očištěnými od outlierů.

První řada grafů vynáší hodnoty v závislosti na pořadovém čísle, v druhém řádku jsou histogramy hodnot. Třetí řádek obsahuje Q-Q diagramy porovnávající kvantily souborů dat s teoretickými kvantily normálního rozdělení.

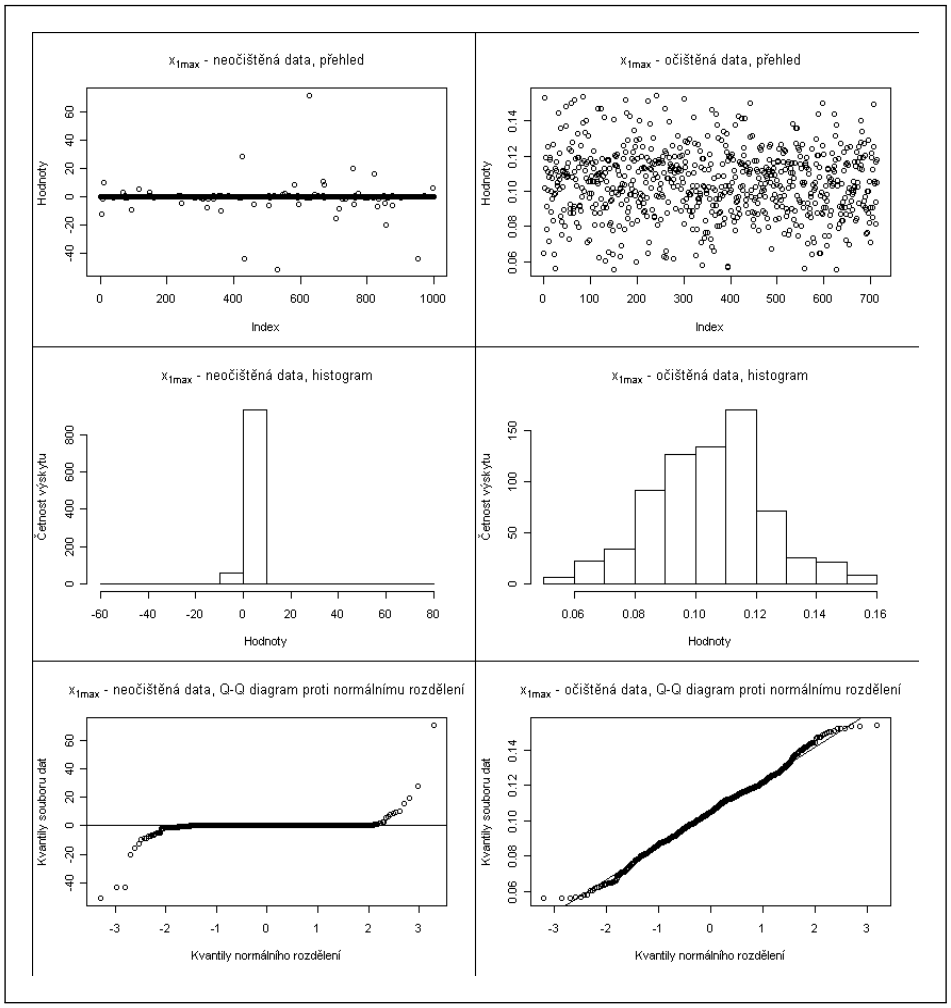
V původních datech se jak u polohy, tak u výšky vyskytují hodnoty pohybující se velice daleko od intervalu, ve kterém bychom čekali výsledek. Zároveň vidíme, že tyto hodnoty silně narušují normalitu rozdělení hodnot. Přítomnost těchto hodnot se dá vysvětlit několika způsoby.

Prvním vysvětlením je, že outliery odpovídají odhadům, kde do vyšetřovaného intervalu $[0{,}06; 0{,}15]$ padlo velice málo hodnot (cca do pěti hodnot). V takovém případě nemusí dát fitování parabolou uspokojivý výsledek. Výsledné parametry, a díky tomu i odhad polohy a výšky maxima, mohou být značně vzdáleny od námi očekávaných. Pokud například hledáme maximálně věrohodný odhad pro hodnoty {0,068, 0,0884, 0,145}, dostaneme $\hat{x}^{ML} = 0{,}1165$ a $\hat{y}^{ML} = -25{,}9372$ ($\hat{a}^{ML} = 33497$, $\hat{b}^{ML} = -7806$).

Druhou možnou příčinou výskytu outlierů je použití numerické metody k maximalizaci věrohodnostní funkce. Pokud má věrohodnostní funkce kromě globálního maxima i další (lokální) maxima, algoritmus může najít právě některé z nich.

Problémy může způsobovat i zastavovací kritérium maximizéru. Pokud by věrohodnostní funkce měla velmi ploché maximum, maximizér by se mohl zastavit daleko od optimálních hodnot. Maximizér by mohl mít problém i s penalizací věrohodnostní funkce, kterou jsme použili.

Díky tomu, že ze souborů odhadů odstraňujeme outliery, omezíme vliv všech těchto rizik. Cenou za to je „umělé“ snížení rozptylu. Je totiž pravděpodobné, že odstraníme



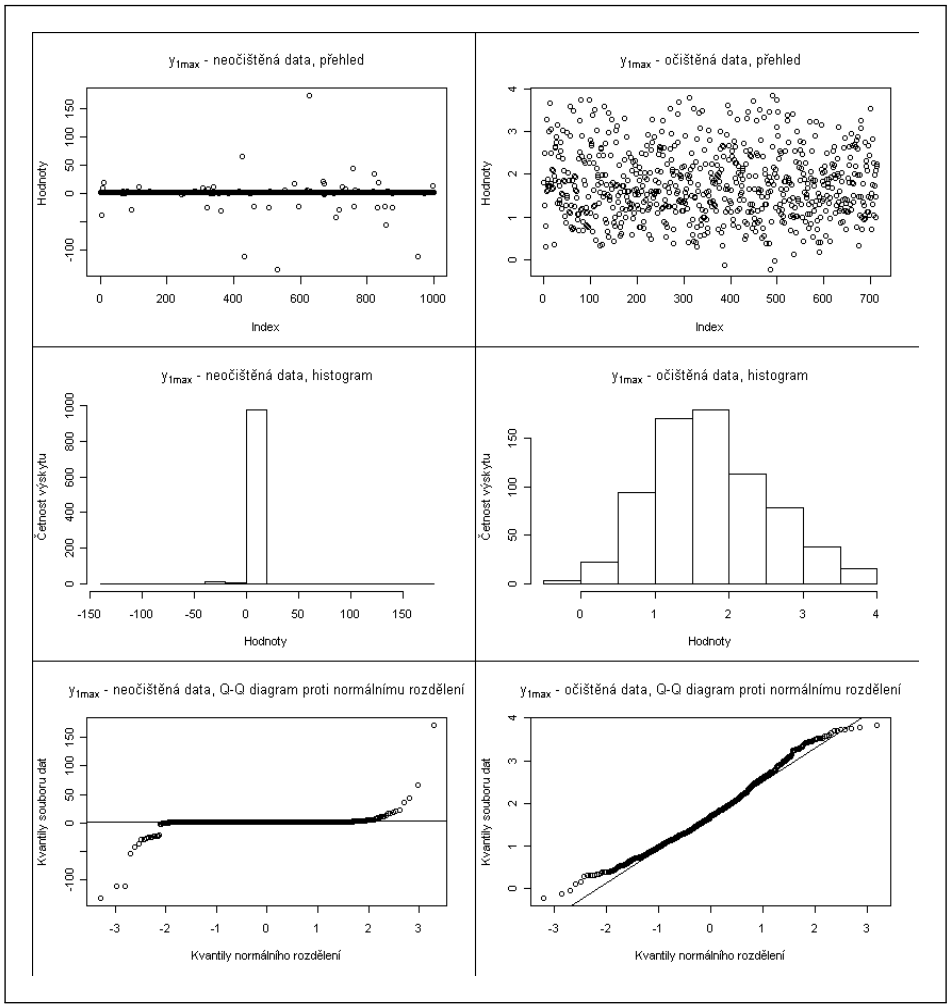
Obrázek 4.8: Přehled hodnot $\hat{x}_{1max}^{ML}$ pro případ $N = 100$, $K = 1000$

i hodnoty, které do souboru „přirozeně“ patří. Druhým negativním efektem může být větší odchylka rozdělení odhadů od normálního rozdělení.

Nyní se podívejme, zda nejsou některé ze získaných odhadů navzájem korelované. Korelační koeficienty pro případ $N = 100$, $K = 1000$ jsou vyneseny v tabulce 4.7.

	l	$\hat{x}^{ML}$	$\hat{y}^{ML}$	$\hat{a}^{ML}$	$\hat{b}^{ML}$	$\hat{c}^{ML}$
l	1	-0,007	-0,025	0,134	-0,133	0,055
$\hat{x}^{ML}$	-0,007	1	0,982	-0,020	0,021	0,015
$\hat{y}^{ML}$	-0,025	0,982	1	-0,039	0,025	-0,007
$\hat{a}^{ML}$	0,134	-0,020	-0,039	1	-0,985	0,935
$\hat{b}^{ML}$	-0,133	0,021	0,025	-0,985	1	0,980
$\hat{c}^{ML}$	0,055	0,015	-0,007	0,935	0,980	1

Tabulka 4.7: Tabulka korelačních koeficientů hodnot získaných metodou maximální věrohodnosti



Obrázek 4.9: Přehled hodnot $\hat{y}_{1max}^{ML}$ pro případ $N = 100$, $K = 1000$

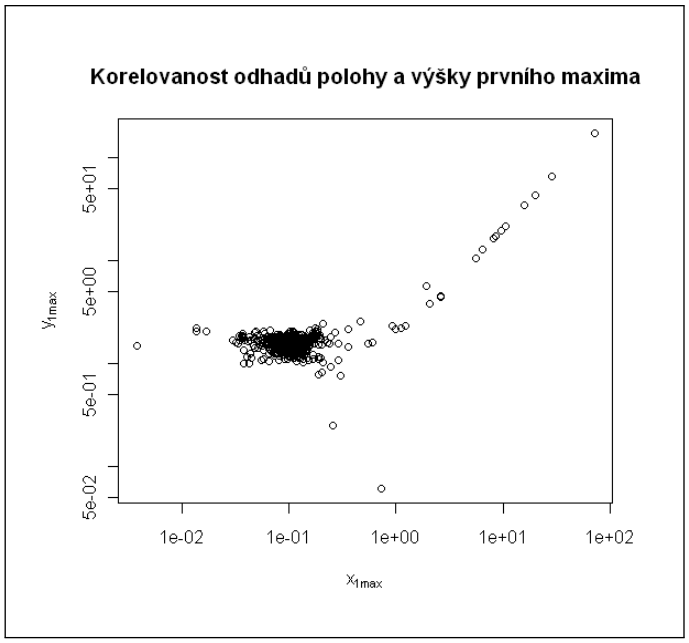
Vidíme, že hodnoty a, b a c jsou navzájem silně korelované. Stejně tak odhady polohy a výšky vykazují silnou korelaci. To naznačuje, že outliery by stačilo hledat buď jen v polohách, nebo jen ve výškách. Žádný ze získaných parametrů není korelovaný s hodnotou věrohodnostní funkce. V tabulce 4.8 jsou pak vyneseny korelační koeficienty, kde dvojice porovnávaných souborů jsou druhé mocniny odchylky od průměru daného parametru a hodnota logaritmu věrohodnostní funkce. Vidíme, že korelace není příliš výrazná. Outliery tedy nelze jednoduše rozpoznat jen z hodnoty věrohodnostní funkce.

Nyní porovnejme přesnost odhadů pomocí maximální věrohodnosti s přesností odhadů vytvořených pomocí histogramů. Jako odhad výšky prvního maxima budeme brát výšku binu, který obsahuje polohu maxima $x_{1max} \doteq 0{,}1036$. Jako chyby uvádíme směrodatnou odchylku z tisíce opakování. Pro šířku binů volíme pravidlo Friedmana a Diaconise (vztah (4.4)). Výsledky shrnuje tabulka 4.9. Vidíme, že histogram výsledky systémově podhodnocuje.

Pokud bychom chtěli použít jádrové odhady, situace bude podobná. Tabulka 4.10

Korelace mezi:	l
$(\hat{x}^{ML} - <\hat{x}^{ML}>)^2$	0,0035
$(\hat{y}^{ML} - <\hat{y}^{ML}>)^2$	0,0024
$(\hat{a}^{ML} - <\hat{a}^{ML}>)^2$	0,1328
$(\hat{b}^{ML} - <\hat{b}^{ML}>)^2$	0,1310
$(\hat{c}^{ML} - <\hat{c}^{ML}>)^2$	0,1341

Tabulka 4.8: Korelační koeficienty mezi hodnotou logaritmu věrohodnostní funkce a odchylkami veličin od průměru pro případ N=100 a $K = 1000$



Obrázek 4.10: Korelace hodnot $\hat{x}^{ML}_{1max} = 0{,}1165$ a $\hat{y}^{ML}_{1max} = 0{,}1165$ pro případ $N = 100$, $K = 1000$

N	100	1000	10000
$\hat{y}^{hist}_{1max}$	$1{,}00 \pm 0{,}18$	$1{,}20 \pm 0{,}11$	$1{,}53 \pm 0{,}06$

Tabulka 4.9: Odhad výšky maxima pomocí histogramů.

obsahuje hodnoty jádrového odhadu v bodě $x = 0{,}1036$ pro různá N. Použito je Epanechnikovo jádro a pravidlo Sheathera a Jonese.

N	100	1000	10000
$\hat{y}^{\text{jádr}}_{1max}$	$0{,}84 \pm 0{,}17$	$1{,}22 \pm 0{,}09$	$1{,}51 \pm 0{,}04$

Tabulka 4.10: Odhad výšky maxima pomocí jádrových odhadů.

Vidíme tedy, že v tomto případě je opravdu výhodnější použít parametrické odhady místo neparametrických. Nejmarkantnější rozdíl je u malých souborů dat, kde je šířka binu či jádra příliš velká na to, aby dobře zachytila situaci v jednom bodě.

Výše zmíněný postup odhadu polohy a výšky maxima pomocí metody maximální věrohodnosti lze samozřejmě dále zlepšovat. Jednou z možností je aproximovat hustotu pravděpodobnosti polynomy vyššího řádu či použít jiné křivky. Úspěšnost lze v tomto případě posuzovat pomocí **Akaikeho informačního kritéria** $AIK = 2k - 2\log(L)$. Kritérium porovnává, jestli použití křivky s více stupni volnosti k dostatečně zlepší věrohodnost odhadu L. Z několika otestovaných prokládaných křivek se vybere ta s nejnižší hodnotou AIK [31].

Druhým směrem je hledání optimálního intervalu, na kterém hustotu pravděpodobnosti aproximujeme. Tímto způsobem by šel snížit počet outlierů na úkor zvýšení rozptylu či naopak. Tyto směry dále rozvíjet nebudeme. Zaměříme se na jiný aspekt metody, který omezuje možnosti metodu aplikovat.

Při testování metody jsme opakovaně generovali náhodné výběry z obecně neznámého rozdělení pravděpodobnosti. Abychom získali rozptyl odhadů, museli jsme provést vyhodnocení všech K výběrů. V praxi je ale často obtížné získat dostatek dat a někdy si musíme vystačit pouze s jednou fotografií. Následující část ukazuje, jak dostupná data využít opakovaně a tím dostat alespoň odhady rozptylu vypočtených hodnot.

4.3 Bootstrap

Metoda bootstrap je poměrně nový[5] statistický koncept. Na rozdíl od odhadů hustoty pravděpodobnosti či statistických testů nefunguje nezávisle, ale pouze modifikuje stávající metody. Využívá se hlavně při nedostatku testovacích dat či pro ověřování přesnosti nějaké statistické metody. Obecně je pak aplikovatelná, kdykoli odhadujeme hodnotu nějaké charakteristiky θ pravděpodobnostního rozdělení náhodného jevu X s obecně neznámým rozdělením na základě náhodného výběru $x_1, \ldots, x_N$ vygenerovaného z tohoto rozdělení.

V této části ukážeme, jak aplikovat metodu bootstrap na metodu maximální věrohodnosti. Zopakujeme celý postup odhadování výšky a polohy prvního maxima hustoty pravděpodobnosti ρ_1 a ukážeme, jak se bootstrap vypořádá s případem, kdy nemůžeme opakovaně vygenerovat dostatek dat ke statistickému zhodnocení přesnosti odhadu. Reálně pak situace bude odpovídat takové situaci, ve které bychom měli k dispozici pouze jednu fotografii zachycující výsledky experimentu a chtěli bychom přesto určit směrodatnou odchylku odhadů pozice a výšky prvního maxima radiální distribuční funkce.

Zopakujme si, jak jsme postupovali v předchozí části. Nejprve jsme vytipovali, v jakém místě se přibližně vyskytuje hledané maximum. Následně jsme hustotu pravděpodobnosti na tomto úseku aproximovali parabolou a hledali jsme, jaké parametry a, b, c nejlépe odpovídají dostupným datům. Z odhadů hodnot těchto koeficientů jsme pak dopočítali odhady $\hat{x}^{ML}_{1max}$ a $\hat{y}^{ML}_{1max}$. To jsme opakovali celkem K-krát, získali jsme tak sady odhadů $\{\hat{x}^{ML}_{1max,i}\}_{i=1}^{K}$, $\{\hat{y}^{ML}_{1max,i}\}_{i=1}^{K}$, ze kterých jsme následně vyloučili extrémní hodnoty.

[5] První článek o bootstrapu pochází z roku 1979 [32].

V tomto okamžiku je vhodné připomenout dvě důležité skutečnosti související s tímto algoritmem. Prvně, pro každý odhad $\hat{x}^{ML}_{1max,i}$ a $\hat{y}^{ML}_{1max,i}$ jsme použili nově vygenerovaná data. Mohli jsme tak postupovat díky tomu, že jsme znali skutečnou hustotu pravděpodobnosti. V naprosté většině případů však hustotu pravděpodobnosti neznáme a naše schopnost vygenerovat data se stejným rozdělením je omezená. Druhou skutečností, úzce související s první, je způsob, jakým jsme získali směrodatné odchylky našich odhadů. Mlčky jsme předpokládali, že $\hat{x}^{ML}_{1max}$ i $\hat{y}^{ML}_{1max}$ jsou náhodné veličiny s neznámým rozdělením. Díky opakovanému počítání těchto odhadů z nově vygenerovaných dat jsme mohli brát výsledné hodnoty jako nezávislé realizace z daného neznámého rozdělení.

Zkusme si nyní celou situaci popsat přesněji. Opakovaně jsme generovali soubory náhodných veličin $\{x_i\}_{i=1}^N$ z náhodného rozdělení s hustotou pravděpodobnosti ρ_1 a s distribuční funkcí, kterou označíme jako F. Pomocí těchto souborů jsme získali sadu odhadů $\theta_1, \theta_2, \dots$. Ze sady odhadů už jsme se mohli pokusit odhadnout rozdělení hodnot odhadů (jehož distribuční funkci označíme jako T, odhad jako $\hat{T}_N$), my jsme se však omezili na odhad rozptylu tohoto rozdělení.

Nyní si představme, že máme k dispozici jen jeden soubor hodnot $\{x_i\}_{i=1}^N$ a neznáme hustotu pravděpodobnosti ani distribuční funkci rozdělení, ze kterého byl vygenerován. Na základě těchto hodnot můžeme hustotu pravděpodobnosti či distribuční funkci pouze odhadnout. Nejjednodušší je použít empirickou distribuční funkci $\hat{F}_N$. Čím větší soubor na počátku máme, tím lépe empirická distribuční funkce aproximuje skutečnou distribuční funkci. Další soubory dat pak budeme generovat z rozdělení $\hat{F}_N$. Takovéto generování odpovídá výběru s vracením z původního souboru dat. Takovéto soubory dat budeme nazývat **bootstrapovými výběry** a budeme je značit $\{x^b_i\}_{i=1}^N$. Použitím těchto souborů můžeme získat odhady veličiny θ. Tyto odhady budeme nazývat **bootstrapovými odhady** a budeme je značit $\hat{\theta}^b$. Konečně na základě souboru bootstrapových odhadů můžeme odhadnout distribuční funkci T. Tento odhad budeme značit $\hat{T}^b_N$.

Nyní se můžeme ptát, jestli se bootstrapový odhad rozdělení $\hat{T}^b_N$ blíží ke stejnému odhadu, při kterém jsme ale generovali hodnoty z rozdělení se skutečnou distribuční funkcí F, tedy k odhadu $\hat{T}_N$. Při splnění vhodných podmínek pro odhady distribučních funkcí platí následující vztahy [33]:

$$\rho_\infty(\hat{T}_N, \hat{T}^b_N) \xrightarrow[N\to\infty]{s.j.} 0\,,$$

$$\rho_\infty(\hat{T}_N, \hat{T}^b_N) \approx \frac{\sqrt{\log\log N}}{\sqrt{N}} \quad \text{pro} \quad N \to \infty\,,$$

přičemž ρ_∞ značí supremovou metriku.

První vztah nám říká, že $\hat{T}^b_N$ je konzistentní odhad $\hat{T}_N$. Druhý nám dává asymptotickou rychlost konvergence. Vidíme, že pro dostatečně velký soubor je rozdíl obou odhadů zanedbatelný. Pokud aplikujeme tyto vztahy při odhadování polohy a výšky prvního maxima, zjistíme, že rozdělení odhadů $\hat{x}^{ML,b}_{1max}$ a $\hat{y}^{ML,b}_{1max}$ bude mít s rostoucím N přibližně stejný rozptyl jako odhady $\hat{x}^{ML}_{1max}$ a $\hat{y}^{ML}_{1max}$. Navíc rozdělení bootstrapových odhadů by se, stejně jako rozdělení „obyčejných“ nebootstrapových odhadů, s rostoucím N mělo blížit k normálnímu rozdělení. Otestujme nyní tyto závěry v praxi.

Metodu bootstrap jsme aplikovali na příklad odhadu polohy a výšky prvního maxima hustoty pravděpodobnosti ρ_1. Veličina N značí velikost souboru vygenerovaného z původního rozdělení, veličina B značí počet bootstrapových výběrů použitých k odhadu veličin $\hat{x}^b_{1max}$ a $\hat{y}^b_{1max}$. Uvedená chyba je pak bootstrapovým odhadem směrodatné odchylky. Stejně jako v předchozí části jsme pomocí Shapirova-Wildova testu zjišťovali, nakolik se rozdělení odhadů blíží normálnímu rozdělení.

N\B	100	1 000	10 000
100	$\hat{x}^{ML,b}_{1max} = 0{,}1091 \pm 0{,}0157$ $\hat{y}^{ML,b}_{1max} = 1{,}1564 \pm 0{,}7404$ počet outlierů: 27	$\hat{x}^{ML,b}_{1max} = 0{,}0965 \pm 0{,}0092$ $\hat{y}^{ML,b}_{1max} = 2{,}7940 \pm 0{,}8055$ počet outlierů: 288	$\hat{x}^{ML,b}_{1max} = 0{,}1056 \pm 0{,}0186$ $\hat{y}^{ML,b}_{1max} = 1{,}7595 \pm 0{,}6038$ počet outlierů: 4703
1 000	$\hat{x}^{ML,b}_{1max} = 0{,}0962 \pm 0{,}0383$ $\hat{y}^{ML,b}_{1max} = 1{,}5517 \pm 0{,}2747$ počet outlierů: 19	$\hat{x}^{ML,b}_{1max} = 0{,}0905 \pm 0{,}0399$ $\hat{y}^{ML,b}_{1max} = 1{,}4244 \pm 0{,}2897$ počet outlierů: 196	$\hat{x}^{ML,b}_{1max} = 0{,}1192 \pm 0{,}0141$ $\hat{y}^{ML,b}_{1max} = 1{,}5782 \pm 0{,}1739$ počet outlierů: 849
10 000	$\hat{x}^{ML,b}_{1max} = 0{,}0981 \pm 0{,}0051$ $\hat{y}^{ML,b}_{1max} = 1{,}5289 \pm 0{,}0459$ počet outlierů: 11	$\hat{x}^{ML,b}_{1max} = 0{,}1018 \pm 0{,}0024$ $\hat{y}^{ML,b}_{1max} = 1{,}5923 \pm 0{,}0525$ počet outlierů: 39	$\hat{x}^{ML,b}_{1max} = 0{,}1080 \pm 0{,}0046$ $\hat{y}^{ML,b}_{1max} = 1{,}5316 \pm 0{,}0535$ počet outlierů: 801

Tabulka 4.11: Souhrn odhadů polohy a výšky prvního maxima metodou maximální věrohodnosti s použitím bootstrapových výběrů dat

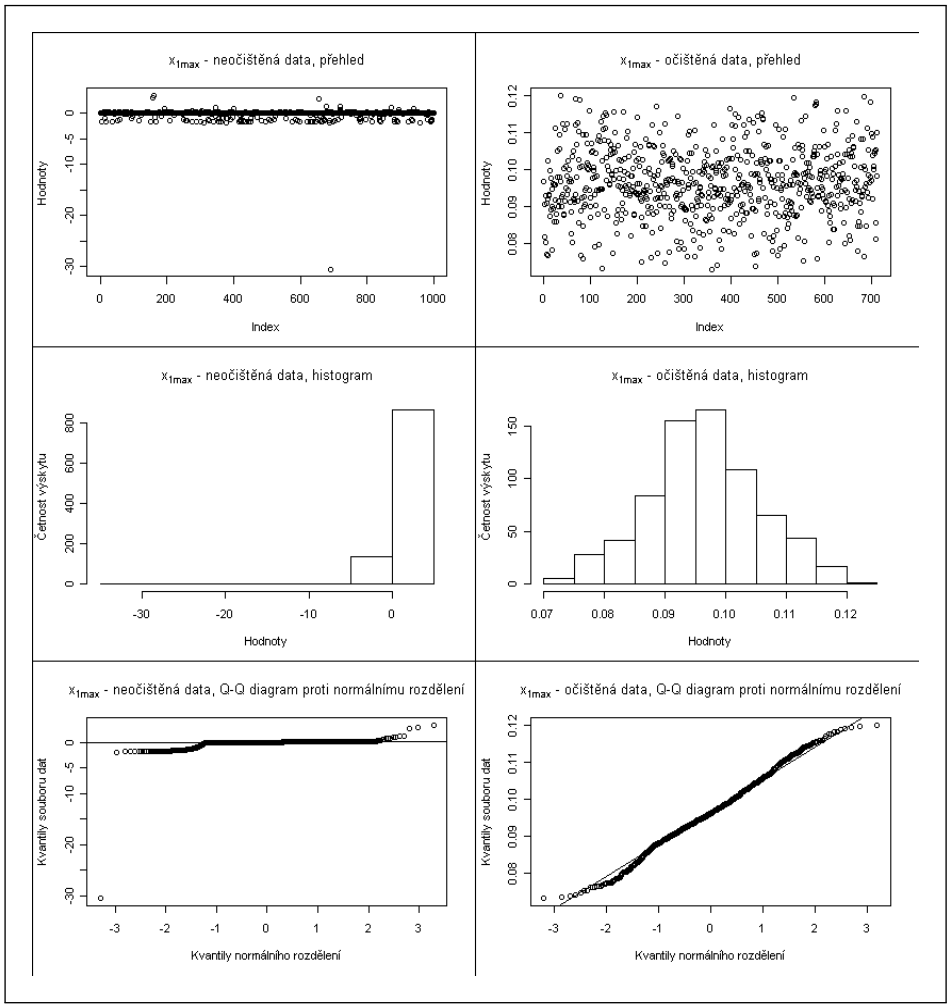
N\B	100	1 000	10 000
100	$p_x^{SW} = 0,915$ $p_y^{SW} = 0,321$	$p_x^{SW} = 8,332 \cdot 10^{-3}$ $p_y^{SW} = 1,238 \cdot 10^{-5}$	$p_x^{SW} = 3,716 \cdot 10^{-12}$ $p_y^{SW} = 1,701 \cdot 10^{-12}$
1 000	$p_x^{SW} = 0,038$ $p_y^{SW} = 0,616$	$p_x^{SW} = 6,170 \cdot 10^{-4}$ $p_y^{SW} = 4,142 \cdot 10^{-3}$	$p_x^{SW} = 1,997 \cdot 10^{-37}$ $p_y^{SW} = 1,516 \cdot 10^{-12}$
10 000	$p_x^{SW} = 0,009$ $p_y^{SW} = 0,772$	$p_x^{SW} = 1,359 \cdot 10^{-1}$ $p_y^{SW} = 2,746 \cdot 10^{-1}$	$p_x^{SW} = 1,238 \cdot 10^{-10}$ $p_y^{SW} = 1,051 \cdot 10^{-6}$

Tabulka 4.12: Přehled p-hodnot Shapirova-Wildova testu pro soubory odhadů polohy a výšky prvního maxima metodou maximální věrohodnosti s využitím bootstrapu

Vidíme, že samotné bootstrapové odhady polohy a maxima jsou dále od skutečných hodnot než v původní metodě. Přesto ve většině případů leží skutečná hodnota v intervalu daném směrodatnou odchylkou. Směrodatná odchylka odhadů zůstává přibližně stejná jako v původní verzi. Na obrázcích 4.11 a 4.12 pak vidíme, že se zásadním způsobem nezměnil ani tvar rozdělení odhadů. Zachována je i míra podobnosti tohoto rozdělení s normálním.

Horší výsledky oproti části 4.2.4 jsou způsobeny tím, že empirická distribuční funkce v některých případech nepřesně aproximovala skutečnou distribuční funkci a tyto nepřesnosti se pak přenesly i do bootstrapových výběrů. Nepřesnosti empirické distribuční funkce naopak příliš neovlivnily tvar distribuce a díky tomu se směrodatné odchylky příliš neliší.

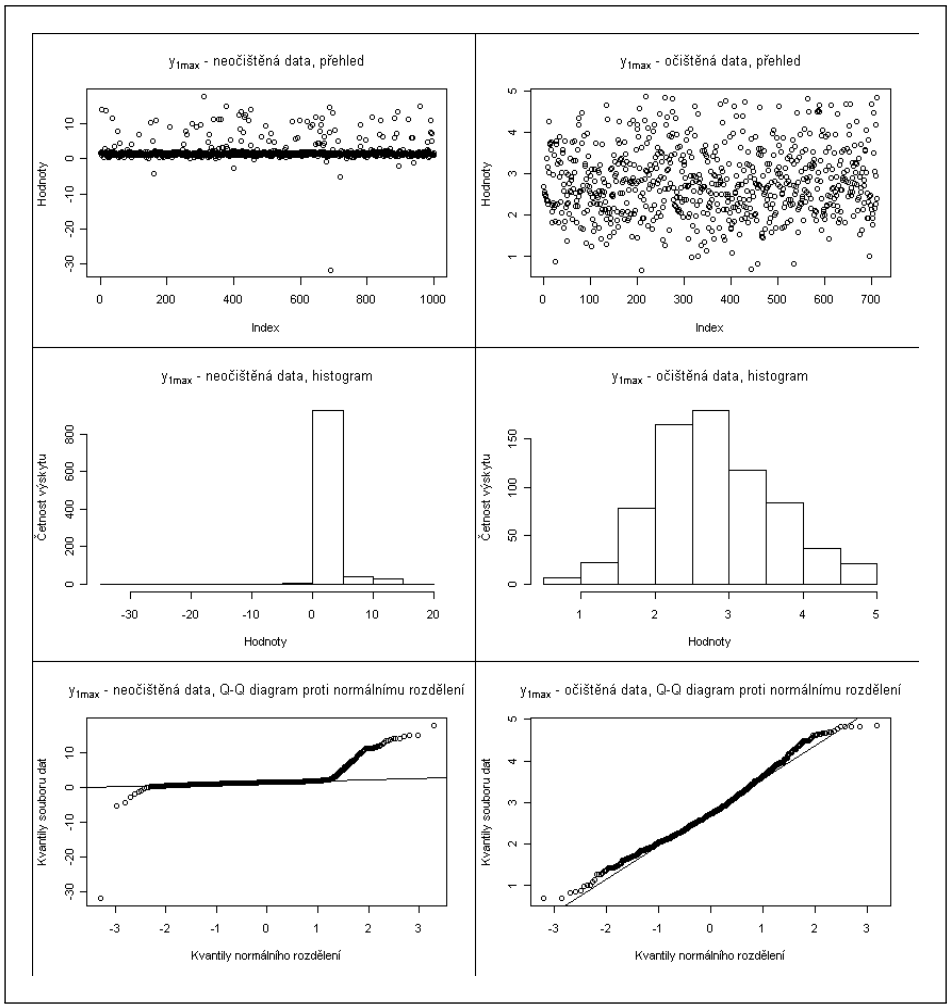
Z výsledků simulace je vidět, že bootstrapové metody lze s úspěchem užívat tam, kde máme omezené množství informací. Musíme si být ale vědomi nepřesností, které



Obrázek 4.11: Přehled hodnot $\hat{x}_{1max}^{ML,b}$ pro případ $N = 100$, $K = 1000$

může bootstrap do výsledku zanést. Další využití bootstrapu se nachází v části 3.2.5, kde je použit k odhadu nepřesností odhadů příznaků morfologických metod.

V této práci jsme využívali metodu bootstrap jen k odhadu směrodatné odchylky některých veličin. S mírnými modifikacemi bychom mohli počítat například konfidenční intervaly daných parametrů, jak ukazuje například monografie [34]. Velké množství dalších aplikací bootstrapu obsahuje kniha [35].



Obrázek 4.12: Přehled hodnot $\hat{y}_{1max}^{ML,b}$ pro případ $N = 100$, $K = 1000$

Kapitola 5

Můj model růstu tenkých vrstev

V kapitole 3 bylo popsáno, jaké vlastnosti očekáváme od dobrých morfologických metod a jejich charakteristik a příznaků. Jednou z uvedených vlastností byla **úplnost**. Tato kapitola bude testovat, nakolik jsou různé charakteristiky úplné - jak přesně lze pouze z charakteristik „zrekonstruovat původní obrázek“. Zrekonstruováním původního obrázku budeme mínit získání obrázku se stejnými vlastnostmi - charakteristikami a příznaky. Zrekonstruovaný obrázek tedy bude takový obrázek, který bychom mohli najít na stejném vzorku jako obrázek původní, či obrázek, který by mohl být vygenerován původním modelem. K tomuto účelu představíme v této kapitole nový model růstu tenkých vrstev.

5.1 Vzájemné vztahy morfologických metod

Ve třetí kapitole jsme rozdělili jednotlivé morfologické metody do několika skupin. Šlo o „integrální informace“, „informace o jednotlivých objektech“ a „informace o rozmístění objektů na ploše“. Dále jsme se zabývali jednotlivými metodami spadajícími do jedné z těchto tří skupin. Už z tohoto rozdělení je vidět, že žádná z charakteristik ani žádný z příznaků uvedených v této práci není úplný a nedokáže popsat charakter obrázku bezezbytku. Pokud bychom však z každé skupiny vybrali jednu metodu, je možné, že vybrané tři metody už by dokázaly popsat obrázek dostatečně dobře.

V úvahu zde přichází i opačný pohled. Představme si, že bychom dokázali z vybraných charakteristik původního obrázku vytvořit obrázek nový, jenž by se shodoval s původním v dalších charakteristikách. Pak by tyto kontrolní charakteristiky obrázku k popisu nebyly potřeba. Kontrolní a původní charakteristiky by nebyly navzájem **nezávislé**.

Jednoduchý případ tří charakteristik/příznaků, které nejsou navzájem nezávislé, jsou počet objektů, rozdělení velikostí objektů a stupeň pokrytí. Pokud bychom znali počet objektů a rozdělení velikostí, mohli bychom z těchto dvou informací spočítat třetí - stupeň pokrytí. Naopak pokud bychom znali počet objektů a stupeň pokrytí, nedokázali bychom získat celkové rozdělení velikostí objektů, ale pouze střední hodnotu velikostí objektů.

Na uvedeném příkladu lze ilustrovat nejen závislost či nezávislost příznaků, ale i to, že uvedené rozdělení morfologických metod do skupin je pouze orientační. Metody

z jedné skupiny nejsou nezávislé na metodách v jiných skupinách. To nám ukázala už dříve analýza příznaků založených na Voronoiově dláždění a Delaunayově triangulaci. Přesto se tohoto rozdělení budeme dále držet.

5.2 Hard-disk jako výchozí bod

V předchozích odstavcích byly uvedeny některé obtíže, s nimiž se musíme vypořádat. Podívejme se nyní, jak danou situaci postihuje model hard-disk. Tento model má tři parametry (velikost plochy zde bereme jako fixní). Vzhledem k tomu, že objekty jsou kruhové, plně je popisuje jejich poloměr. Difúzní parametr vnáší informace o vzájemném postavení objektů, zároveň také reguluje maximální počet objektů. Poměr zaplnění pak primárně reguluje počet objektů na ploše, zároveň ale ovlivňuje rozložení objektů v ploše, jak bylo ukázáno v sekci 3.2.5.

Vidíme, že parametry hard-disk modelu neodpovídají třem uvedeným skupinám metod. Prvním cílem je tedy model modifikovat tak, aby toto splňoval. Druhým nedostatkem hard-disk modelu je nedostatečný popis samotných objektů. V základní verzi modelu jsou všechny objekty identické[1].

První cíl můžeme splnit tak, že počet objektů budeme brát jako úměrný velikosti plochy Ω a místo difúzního parametru použijeme jinou, vhodnější charakteristiku. Problém s nedostatečným popisem pak můžeme jednoduše vyřešit tím, že budeme pro každý objekt generovat pokaždé jinou velikost (případně i jiný tvar).

5.3 Popis modelu

Nyní přejděme k technickému popisu modelu. Jak už bylo řečeno, cílem modelu je zrekonstruovat obrázek za pomoci některých jeho charakteristik. V této části popíšeme, jak může fungovat model, který jako výchozí informace bude používat počet objektů, rozdělení velikostí objektů a radiální distribuční funkci.

Základním úkolem modelu je postupně umísťovat objekty na plochu podle určitých pravidel. Stejně jako v hard-disk modelu řekneme, že objekt leží na ploše Ω, pokud tam bude ležet jeho těžiště.

Než přistoupíme k samotnému umísťování, vygenerujeme si velikosti objektů (pro jednoduchost budeme pracovat pouze s kruhovými objekty), které budeme na plochu vkládat. Velikosti objektů budeme generovat na základě odhadu rozdělení velikostí v původním obrázku. Jako vhodný odhad můžeme vzít histogram či jádrový odhad. Další alternativou by bylo vzít jako velikosti objektů, které budeme vkládat na plochu, bootstrapový výběr z původního souboru velikostí. Počet objektů které, budeme vkládat, zůstane stejný jako na původního obrázku.

Počet objektů z původního obrázku zachováme. Výsledek modelu by pak měl mít stejný stupeň pokrytí jako původní obrázek. To bude prvním testem funkčnosti modelu.

[1] V literatuře [5] lze najít i pokročilejší varianty HD modelu, kde velikost difuzniho parametru D je ovlivňována právě velikostí objektu.

Poslední charakteristikou, kterou použijeme jako zdrojové informaci pro model, bude radiální distribuční funkce. Jako odhad této charakteristiky opět můžeme použít histogram či metodu jádrových odhadů. Jak vyplývá z následujícího popisu, zachování této veličiny nemusí být automatické. Míra zachování tvaru radiální distribuční funkce bude druhým testem modelu. Třetím testem bude porovnání šesti příznaků použitých v analýze obrázku 1.3 v sekci 3.2.5, tedy příznaků označených jako V_t^1, V_{diskr}^1, V_t^2, V_{diskr}^2, D_t^4 a D_{circ}^4.

Nyní už k samotnému umísťování objektů na plochu. Pro každý objekt vygenerujeme náhodně pozici. Prověříme, jestli se do vzdálenosti r_{max} vyskytuje nějaký objekt. Pokud ne, objekt umístíme. Pokud existuje právě jeden takový objekt, umístíme jej s pravděpodobností úměrnou hodnotě radiální distribuční funkce pro tuto vzdálenost. Pokud se ve sledovaném okolí vyskytuje více objektů, příslušné pravděpodobnosti násobíme. Vztah mezi radiální distribuční funkcí a pravděpodobností umístění volíme tak, aby pro maximální hodnotu ρ_{RDF} byla pravděpodobnost rovna jedné.

5.4 Výsledky modelu

V případě radiální distribuční funkce pro nebodové objekty máme tři možnosti. Varianta ρ_{RDF_1} pracuje s objekty jako bodovými. Místo celého objektu používá jen jeho těžiště. Druhá varianta je určena pro diskrétní obrázky, tu zde využívat nebudeme. Třetí varianta ρ_{RDF_3} zohledňuje velikost objektů. Model jsme vyzkoušeli pro první i třetí variantu radiální distribuční funkce. Tyto varianty modelu budeme označovat jako model A a model B. Jako původní obrázek jsme využili výsledek MC modelu popsaného v sekci 2.3 pro napařovací rychlost 10^2 ML/s a tloušťku vrstvy 5 ML.

V tabulce 5.1 vidíme výsledky prvního testu. Výsledky testu nezávisí na variantě použité radiální distribuční funkce. Jako chyba hodnoty u nového stupně pokrytí je uvedena směrodatná odchylka pro 1000 opakování.

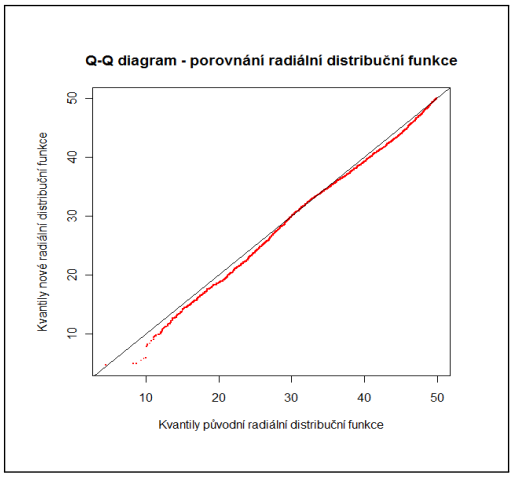
Původní stupeň pokrytí	0,4340
Nový stupeň pokrytí	$0{,}4334 \pm 0{,}0051$

Tabulka 5.1: Výsledky prvního testu - porovnání stupně zaplnění obrázků

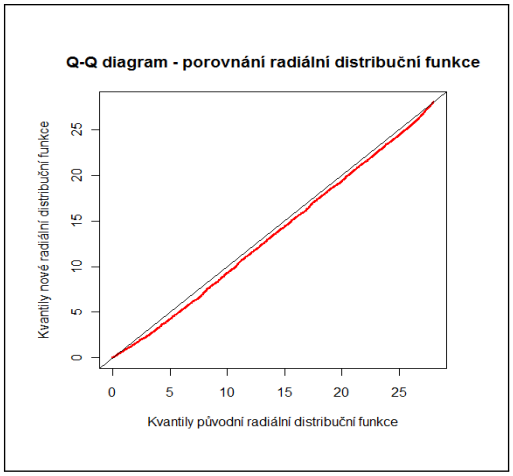
U dalších testů jsou vždy uvedeny výsledky pouze z jedné realizace modelu (pro obě varianty modelu). Uvedené nepřesnosti jsou bootstrapovými odhady směrodatné odchylky příslušných veličin založené na vyhodnocení 1000 bootstrapových výběrů.

Druhým testem je porovnání původní a nové radiální distribuční funkce. Na obrázcích 5.1 a 5.2 můžeme vidět Q-Q diagram porovnávající původní a novou radiální distribuční funkci. V prvním případě je vidět, že shoda je zde velice dobrá. Kolmogorovův-Smirnovův test nám však říká, že se charakteristiky přesto liší (p-hodnota = 0,0027). V druhém případě je shoda mírně horší, p-hodnota KS testu je 0,0004).

Další testy spočívají v porovnání příznaků založených na Voronoiově dláždění a Delaunayově triangulaci. Porovnání hodnot příznaků je v tabulce 5.2, na obrázcích 5.3 a 5.4 jsou vykresleny Q-Q diagramy porovnávající klíčové veličiny příznaků. Vidíme,



Obrázek 5.1: Q-Q diagram pro původní a novou radiální distribuční funkci ρ_{RDF_1}



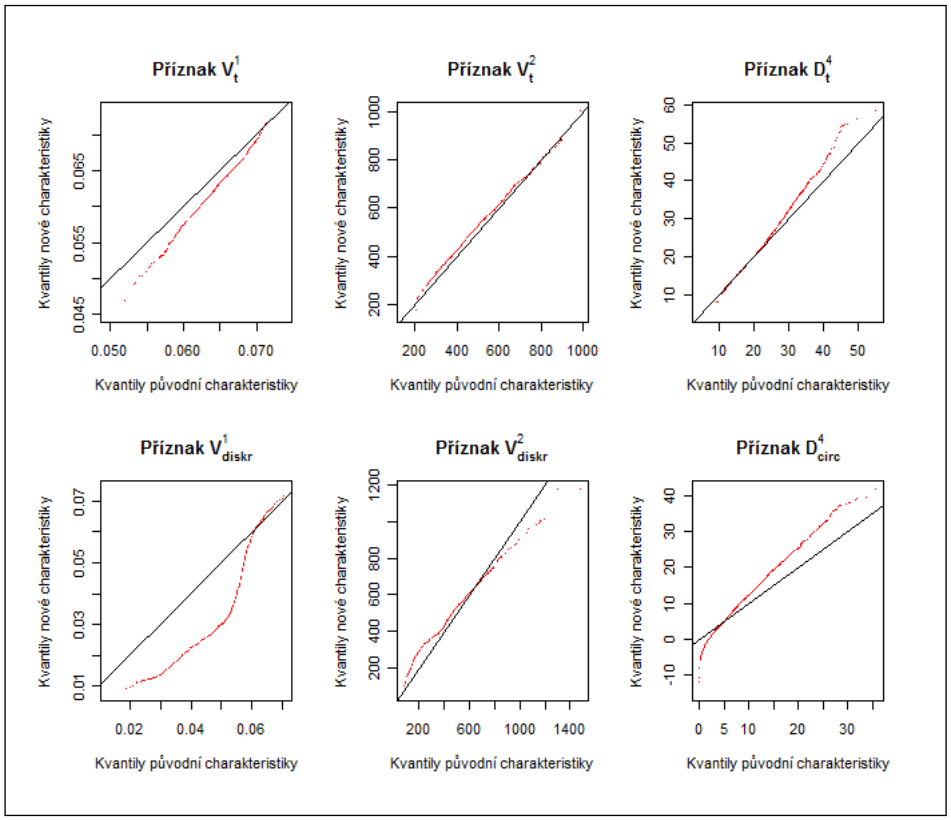
Obrázek 5.2: Q-Q diagram pro původní a novou radiální distribuční funkci ρ_{RDF_3}

že pro první variantu radiální distribuční funkce je shoda příznaků i průběh rozdělení klíčových veličin velmi špatný. Model zachoval nejlépe příznak V_t^2.

Výsledky pro druhý test jsou lepší. Model velmi dobře zachoval příznaky V_{diskr}^1 a D_{circ}^4. Právě u těchto příznaků byla shoda s původními u modelu A nejhorší. U ostatních příznaků je podobnost průběhů původních a nových veličin přibližně shodná.

Na závěr se podívejme na obrázky vytvořené modelem. Původní data jsou na obrázku 5.5, výsledek modelu A je na obrázku 5.6, výsledek modelu B můžeme vidět na obrázku 5.7.

Výsledek první verze modelu je s originálem nesrovnatelný. Největším rozdílem je vzájemné překrývání objektů na novém obrázku. Na vině je radiální distribuční funkce ρ_{RDF_1}, jež nezahrnuje informace o velikostech objektů a vzdálenosti objektů měří mezi jejich těžišti. Tím separuje informace o objektech od informací o rozložení objektů. Při zpětné rekonstrukci pak neexistuje klíč, jak propojit velikosti objektů se



Obrázek 5.3: Porovnání příznaků založených na Voronoiově dláždění či Delaunayově triangulaci pro model používající ρ_{RDF_1}

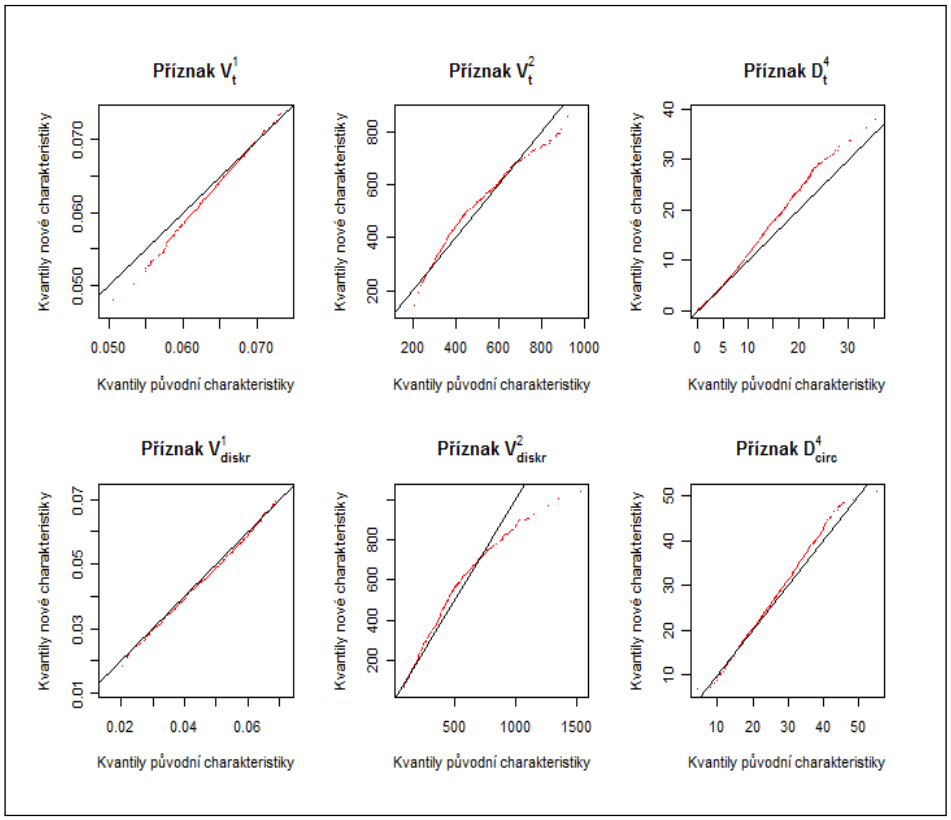
	Původní hodnota	Nová hodnota model A	Nová hodnota model B
V_t^1	0,964	$1{,}087 \pm 0{,}013$	$1{,}058 \pm 0{,}013$
V_{diskr}^1	2,110	$3{,}447 \pm 0{,}045$	$2{,}143 \pm 0{,}029$
V_t^2	0,0236	$0{,}0223 \pm 0{,}0002$	$0{,}0216 \pm 0{,}0002$
V_{diskr}^2	0,0296	$0{,}0257 \pm 0{,}0003$	$0{,}0270 \pm 0{,}0004$
D_t^4	0,105	$0{,}111 \pm 0{,}001$	$0{,}107 \pm 0{,}001$
D_{circ}^4	0,283	$0{,}307 \pm 0{,}003$	$0{,}284 \pm 0{,}002$

Tabulka 5.2: Výsledky třetího testu - porovnání příznaků založených na Voronoiově dláždění a Delaunayově triangulaci

vzdálenostmi jejich sousedů. Nejlepší výsledek, který by mohl model A vytvořit, by odpovídal originálnímu obrázku s promíchanými velikostmi objektů.

Výsledek druhého modelu se k originálu blíží. Přesto však lze najít výrazné rozdíly. Prvním je netypické hromadění objektů u okraje obrázku. Tohoto efektu by se dalo zbavit, pokud by se do modelu zahrnuly periodické okrajové podmínky. Nyní se objekty hromadí u okrajů jednoduše proto, že v jejich okolí je méně objektů. Pravděpodobnost umístění se počítá jako součin pravděpodobností pro vzdáleností jednotlivých objektů. Pravděpodobnost pro pozici na okraji bude součinem menšího počtu členů.

Druhého rozdílu si můžeme všimnout na originálním obrázku. Kolem velkých ob-

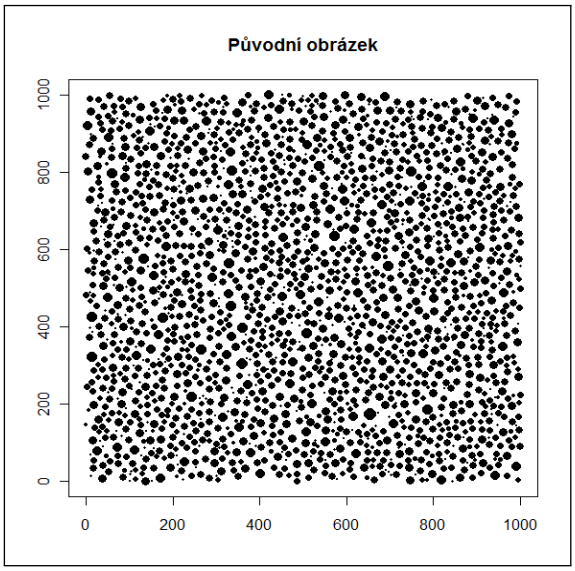


Obrázek 5.4: Porovnání příznaků založených na Voronoiově dláždění či Delaunayově triangulaci pro model používající ρ_{RDF_3}

jektů je často volný prostor. Ten vznikl pohlcením menších objektů. Tento jev z MC modelu odpovídá i skutečným experimentálním datům. Tento efekt se ve výsledku modelu nevyskytuje. Radiální distribuční funkce sice zohledňuje velikost objektů, neobsahuje však informaci o jejich okolí. První maximum odpovídající průměrné vzdálenosti přirozených sousedů není rozlišeno z hlediska velikosti objektů. Z tohoto pohledu by model mohl dosahovat dobrých výsledků, pokud by zpracovával a uchovával radiální distribuční funkce pro různé kombinace velikostí objektů.

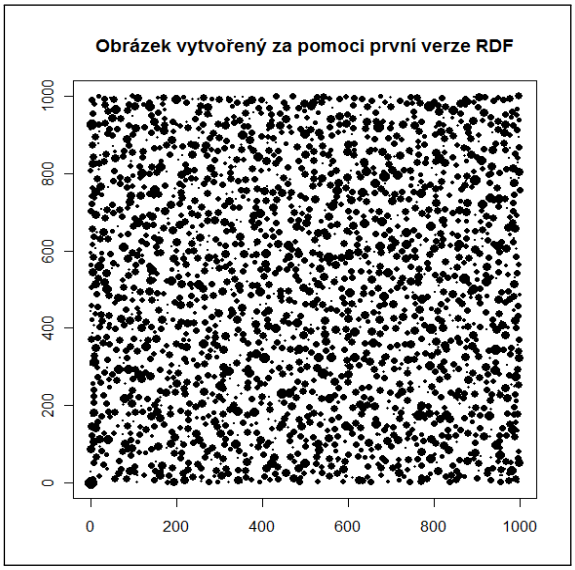
Z výše uvedených výsledků modelu si můžeme udělat závěr, že počet objektů, rozdělení velikostí a radiální distribuční funkce dohromady neobsahují všechny informace o obrázku. Pokud tedy chceme obrázek rekonstruovat, musíme použít více příznaků či charakteristik nebo použít jiné kombinace.

Na závěr uvedeme pár poznámek týkajících se technických aspektů modelu. Rychlost rekonstrukce obrázku závisí na několika veličinách. První z nich je zřejmá, jde o počet vkládaných objektů. Rychlost modelu závisí též na tvaru radiální distribuční funkce a na parametru r_{max} z definice ρ_{RDF}. Čím větší je parametr r_{max}, tím větší okolí potenciální nové pozice objektu je bráno v úvahu. Dále může ovlivnit rychlosti objektu samotný tvar radiální distribuční funkce. Čím je první maximu vyšší, tím je původní obrázek uspořádanější. Tím je ale nepravděpodobnější umístění objektu do vzdálenosti. To zvyšuje průměrný počet neúspěšných pokusů o umístění objektu.

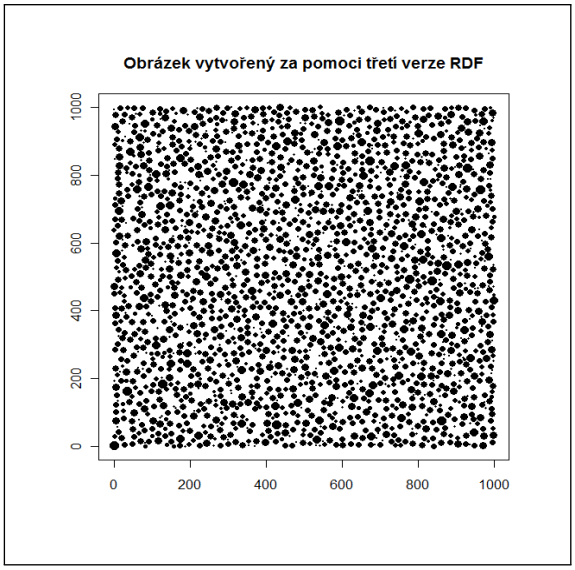


Obrázek 5.5: Originální obrázek, jehož radiální distribuční funkci model používal

Obrázek, který byl modelem rekonstruován, obsahoval kolem 2 000 objektů. V závislosti na volbě parametru r_{max} trvala jedna rekonstrukce řádově několik minut až několik desítek minut.



Obrázek 5.6: Výsledek modelu při použití ρ_{RDF_1}



Obrázek 5.7: Výsledek modelu při použití ρ_{RDF_1}

Kapitola 6

Shrnutí původních výsledků

Nyní přichází čas shrnout všechny originální výsledky uvedené v této práci. V tomto shrnutí budeme postupovat chronologicky, výsledky budeme uvádět v takovém pořadí, v jakém se v práci objevily.

První kapitola je pouze rešeršní a neobsahuje žádné původní výsledky. V druhé kapitole je originálním výsledkem algoritmus, který zjišťuje, jaké největší množství objektů lze vložit na plochu objektů v modelu hard-disk. Tento algoritmus zároveň zkracuje dobu potřebnou k vygenerování modelu s poměrem zaplnění rovným jedné.

V třetí kapitole jsme si představili novou variantu radiální distribuční funkce, jež vykazuje lepší výsledky než původní varianta, na datech s plošnými objekty. Zároveň jsme navrhli modifikaci algoritmu výpočtu původní verze radiální distribuční funkce s cílem lépe odhadnou průběh radiální distribuční funkce z dostupných dat.

U morfologických metod jsme představili příznaky založené na Delaunayově triangulaci (příznaky založené na Voronoiově dláždění se v literatuře používají). Vybrané příznaky jsme pak modifikovali tak, aby lépe charakterizovaly rozložení objektů v případě, že objekty jsou reprezentovány množinami pixelů na digitální fotografii.

Čtvrtá kapitola obsahovala kromě rešeršní části věnované statistickým metodám i ukázky jejich aplikace na výpočet příznaků morfologických metod. Nový je především způsob odhadu velikosti maxima funkce pomocí metody maximální věrohodnosti a odhad nepřesnosti tohoto určení pomocí metody bootstrap.

V páté kapitole je představen nový model růstu tenkých vrstev, jež má za cíl vytvořit z původní fotografie růstu tenkých vrstev fotografii novou, tak aby její morfologické charakteristiky byly co nejpodobnější fotografii původní. Pomocí tohoto modelu lze vyšetřovat nezávislost a úplnost různých kombinací morfologických metod.

Seznam obrázků

Seznam tabulek

Literatura

[1] Eckertová L.: **Physics of Thin Films**, SNTL, Praha, 1986.

[2] Hrach R.: **Počítačová fyzika I**, skripta PF UJEP, Ústí nad Labem 2003.

[3] Hrach R.: **Počítačová fyzika II**, skripta PF UJEP, Ústí nad Labem 2003.

[4] Rozsíval V.: **Morfologie růstu tenkých kovových vrstev na dielektrických podložkách**, bakalářská práce, Univerzita Jana Evangelisty Purkyně v Ústí nad Labem, 2005.

[5] Hrach R. et al.: **Morphological analysis of discontinuous and semicontinuous metal films**, Thin Solid Films, Volume 317, Issues 1-2, 1 April 1998, 39-42.

[6] Hrach R., Sobotka M.: **Methods of mathematical morphology in spatial analysis of island metal films**, International Journal of Electronics, Volume 69, Issue 1, 1990, 49-54.

[7] Okabe A. et al.: **Spatial Tessellation: Concepts and Applications of Voronoi Diagrams**, second edition, J. Wiley and Sons, 2000.

[8] Kostern M.: **Metody matematické morfologie a integrální transformace ve fyzice tenkých vrstev**, dizertační práce, MFF UK, Praha 2007.

[9] Škvor J.: **Použití teorie perkolace ve fyzice tenkých vrstev**, Diplomová práce, Univerzita Jana Evangelisty Purkyně v Ústí nad Labem, 2005.

[10] Hrach R. et al.: **Computer simulation of semicontinuous and continuous metal film morphology**, Vacuum, Volume 50, Issues 3-4, 1 July 1998, Pages 289-292.

[11] Klepeis J.L. et al.: **Long-timescale molecular dynamics simulations of protein structure and function**, Current Opinion in Structural Biology, Volume 19, Issue 2, Theory and simulation / Macromolecular assemblages, April 2009, Pages 120-127.

[12] Hrach R. Novotný D., Hrubý V.: **Study of initial stages of thin film growth by means of atomistic computer simulation and image analysis**, Vacuum (2011), v tisku.

[13] Kuriščák P.: **Studium počátečních fází růstu kovových vrstev postupy počítačové fyziky**, bakalářská práce, Katedra fyziky povrchů a plasmatu, MFF UK, Praha 2009.

[14] R Development Core Team (2010): **R: A language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org/`.

[15] Marsaglia G., Tsang W.W., Wang J.: **Evaluating Kolmogorov's Distibution**, Journal of the American Statistical Association 69 (347), 730-737.

[16] Birnbaum, Z.W.: **Numerical tabulation of the distribution of Kolmogorov's statistic for finite sample size**, J. Amer. Statist. Assoc. 47 (1952), 425-441.

[17] Royston P.: **Remark AS R94: A remark on Algorithm AS 181: The W test for normality**, Applied Statistics, 44, 547-551.

[18] Hyde, J.: **Testing survival with incomplete observations.** Biostatistics Casebook. R.G. Miller, B. Efron, B.W. Brown and L.E. Moses (editors), 31–46. John Wiley, 1980.

[19] Sturges H.A.: **The Choice of a Class Interval**, Journal of the American Statistical Association, Vol. 21, No. 153 (Mar., 1926), pp. 65-66.

[20] Scott D.W.: **Sturges' rule**, JohnWiley&Sons, Inc. WIREsCompStat 2009 1 303–306.

[21] Scott D.W.: **Multivariate Density Estimation, Theory, Practice, and Visualization**, NewYork: JohnWiley&Sons; 1992.

[22] Scott D.W.: **On Optimal and Data-Based Histograms**, Biometrika, Vol. 66, No. 3 (Dec., 1979), pp. 605-610.

[23] Freedman D., Diaconis P.: **On the Histogram as a Density Estimator: L 2 Theory**, Z. Wahrscheinlichkeitstheorie verw. Gebiete 57, 453-476 (1981).

[24] Terrell G.R., Scott D.W.: **Oversmoothed Nonparametric Density Estimates**, Journal of the American Statistical Association, Vol. 80, No. 389 (Mar., 1985), pp. 209-214.

[25] Terrell G.R.: **The Maximal Smoothing Principle in Density Estimation**, Journal of the American Statistical Association, Vol. 85, No. 410 (Jun., 1990), pp. 470-477.

[26] Scott D.W., Terrell G.R.: **Biased and Unbiased Cross-Validation in Density Estimation**, Journal of the American Statistical Association, Vol. 82, No. 400 (Dec., 1987), pp. 1131-1146.

[27] Sheather S.J., Jones M.C.: **A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation**, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 53, No. 3 (1991), pp. 683-690.

[28] Severini T.A.: **Likelihood Methods in Statistics**, Oxford University Press, Oxford, 2000.

[29] Broyden, C. G.: **The convergence of a class of double-rank minimization algorithms**, Journal of the Institute of Mathematics and Its Applications 6: 76–90, 1970.

[30] Wilcox, R.R.: **Introduction to robust estimation and hypothesis testing**, Academic Press, 2005.

[31] Wikipedia contributors: **Akaike information criterion**, Wikipedia, The Free Encyclopedia, `http://en.wikipedia.org/w/index.php?title=Akaike_information_criterion&oldid=425200963`

[32] Efron, B.: **Bootstrap Methods: Another Look at the Jackknife**, The Annals of Statistics 7 (1): 1–26 (1979).

[33] Prášková Z.: **Metoda bootstrap**, sborník konfernce ROBUST 2004. Dostupné na internetu na `www.statspol.cz/robust/robust2004/praskova.pdf`.

[34] Efron, B., Tibshirani, R.: **An Introduction to the Bootstrap**, Boca Raton, FL: Chapman & Hall/CRC (1993).

[35] Davison, A. C., Hinkley, D.: **Bootstrap Methods and their Application.** (2006). Bootstrap Methods and their Application (8th ed.). Cambridge: Cambridge Series in Statistical and Probabilistic Mathematics.

Rejstřík