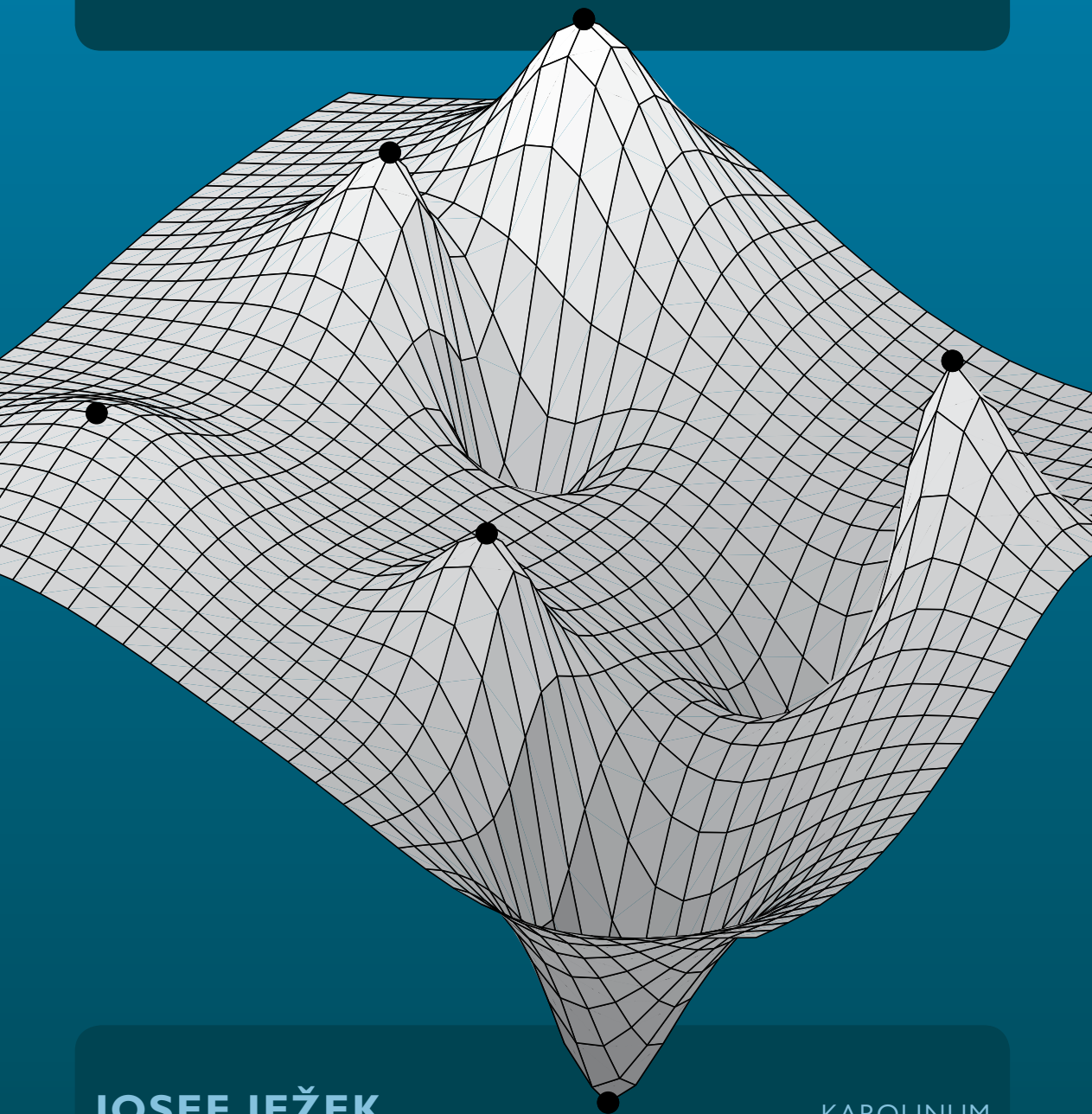


GEOSTATISTIKA A PROSTOROVÁ INTERPOLACE



JOSEF JEŽEK

KAROLINUM

Geostatistika a prostorová interpolace

Josef Ježek

Recenzovali:

doc. Dr. Ing. Jiří Horák, CSc.

doc. RNDr. Karel Zvára, CSc.

Vydala Univerzita Karlova v Praze

Nakladatelství Karolinum

Obálka Jan Šerých

Sazba DTP Nakladatelství Karolinum

Vydání první

© Univerzita Karlova v Praze, 2015

© Josef Ježek, 2015

ISBN 978-80-246-3076-2

ISBN 978-80-246-3104-2 (online : pdf)



Univerzita Karlova v Praze
Nakladatelství Karolinum 2016

www.karolinum.cz
ebooks@karolinum.cz

OBSAH

Předmluva	9
1 Úvod	11
Trochu historie	11
Prostorová interpolace	13
2 Metoda IDW	17
Vlastnosti metody IDW	19
Přesná interpolace vs. shlazení	23
Omezení okolí a anizotropie	25
Vliv konfigurace bodů	27
IDW a jiné metody	28
Cross-validation	29
Interpolace plošných dat	31
Shrnutí metody IDW	36
3 Metoda RBF	39
Interpolace multikvadrikou	39
Obecný tvar interpolace RBF	40
Vliv trendu	42
Základní typy báзовých funkcí	44
Interpolace plošných dat	45
Aproximační varianta RBF	47
4 Polynomický trend	49
Globální trend	50
Lokální trend	52
5 Variogram a kovariance	57
Variogram mrak	57
Empirický variogram	59
Teoretický variogram	63
Meze použitelnosti variogramu	64
Anizotropie	66
Kovariance	66
Prostorová kovariance	67
Vztah kovariance a variogramu	69
6 Prostorová proměnná	71
Náhodná funkce	71

Stacionarita	75
Statistická inference	77
Příklady náhodných funkcí	79
Splnění předpokladů stacionarity	85
Koncepční poznámky	86
7 Modely variogramu	89
Základní typy variogramu	91
Kombinace variogramů	96
Příklady simulací	96
Shrnutí vztahu variogram – prostorová proměnná	101
Geometrická anizotropie	102
Proložení teoretického variogramu	103
8 Kriging	107
Základní kriging	108
Chyba krigingu	112
Počítačová implementace	114
Vlastnosti krigingu	115
Mikrovariabilita vs. chyby měření	119
9 Vypracovaný příklad	125
Analýza dat	126
Strukturní analýza	129
Kriging	133
Analýza trendu	139
Anizotropie	141
Hustší síť	147
10 Duální formulace krigingu	149
11 Kriging střední hodnoty	153
12 Jednoduchý kriging	157
13 Univerzální kriging	159
Určení trendu a jeho koeficientů	162
Určení variogramu stacionární složky	163
Kriging trendu plus jednoduchý kriging	165
Souvislost s metodou zobecněných nejmenších čtverců	166
14 Kriging agregovaných dat	167

15 Porovnání použitých metod	173
16 Náhodné simulace	177
Bibliografie	181
Programy	183
Rejstřík	197

Předmluva

Na otázku, co je *geostatistika*, odpoví mnozí, že je to asi jakýsi souhrn statistických metod používaných v geovědách. Patrně by tomu tak i bylo, nebýt francouzského matematika George Matherona, který na začátku 60. let minulého století použil název *geostatistika* pro označení specifické metody zpracování prostorových dat. Matheronova metoda měla původně sloužit k odhadu vydatnosti ložisek surovin. S jejím rozvojem se však postupně objevily aplikace prakticky ve všech oborech, které pracují s prostorovými daty. Geostatistika proto dnes patří mezi základní metody statistiky prostorových jevů.

Tento text má za cíl vysvětlení základů geostatistiky se zaměřením na prostorovou interpolaci. Vychází z mých přednášek na Přírodovědecké fakultě Univerzity Karlovy v Praze. Inspirací k nim mi byly opakované pobyty v Centre de Géostatistique ve Fontainebleau. Za tuto možnost bych zde chtěl poděkovat.

Vedle geostatistických postupů jsou vysvětleny i vybrané interpolační metody, které s geostatistikou, resp. s jejími aplikacemi bezprostředně souvisejí. V centru zájmu je případ, kdy data jsou zadána v ploše, v nepravidelné síti bodů.

Pokusil jsem se o výklad srozumitelný nejen studentům, ale též uživatelům programů s geostatistikou mimo akademickou obec. Pokud se někde objeví náznak

nějakého matematického postupu, například odvození rovnic krigingu, jde spíše o ukázkou, jakou cestou se k výsledku dospěje, než o exaktní odvození. Jelikož se většinou jedná o základní fakta, neuvádím, až na několik výjimek, citace. Vyčerpávající přehled geostatistické literatury je obsažen v monografiích uvedených na konci knihy. V textu se místy používají anglické názvy a zkratky, neboť relevantní česká terminologie není ustálená a probírané metody jsou dostupné prostřednictvím programů komunikujících anglicky. Neuškodí si příslušnou terminologii osvojit.

Výklad je ilustrován příklady naprogramovanými v Matlabu. Vybrané ukázky programů jsou uvedeny na konci knihy. Sestavení programů pro jednoduché případy nebo jejich modifikace jsou důležité pro osvojení metod. V programování není možno mlžit a kroky algoritmu musejí být konkrétní a jednoznačné. Když studenti některé programy sami vypracují, zjistí, že geostatistika není tak složitá, jak se na první pohled zdá, a že na to mají.

V textu se jistě, přes veškerou moji snahu, objeví chyby. Budu čtenářům vděčný za upozornění na ně i za jakékoliv další připomínky, poslané např. na můj email jezek@natur.cuni.cz. Děkuji všem, kteří byli ochotni text nebo jeho části přečíst a poukázat na nedostatky. Zejména to byli recenzenti Jiří Horák a Karel Zvára, dále kolega Stanislav Bek a též moje žena Ajka. V neposlední řadě si zaslouží můj dík Hans Wackernagel, který mi připomněl, že bych měl napsat o geostatistice něco česky.

Josef Ježek, 13. 2. 2015

/1/ Úvod

TROCHU HISTORIE

V padesátých letech minulého století probíhala po celém světě intenzivní těžba surovin a rozvíjely se metody vyhodnocení ložisek. Důležitou roli hrála jihoafrická škola a zejména důlní inženýr Daniel Krige. Základem jeho metody odhadu bylo použití vážených průměrů. Problémem bylo, jak objektivizovat a optimalizovat volbu vah. V té době přišel Georges Matheron. Matheron byl absolventem elitních francouzských technik, Grandes Écoles, a jeho hluboké vzdělání v matematice a statistice spolu se znalostí problematiky těžby surovin se projevilo ve způsobu, jakým přistoupil k řešení problému. Vytvořil a systematicky rozvíjel metodu, jejíž hlavní postup nazval na počest Daniela Krige *kriging*. Stal se profesorem na École Normale Supérieure des Mines de Paris a ředitelem výzkumného ústavu Centre de Géostatistique zřízeného touto školou ve Fontainebleau. Geostatistika měla hned od počátku mnoho přívrženců, ale i odpůrců.

Matheron použil pravděpodobnostní popis, což způsobilo problémy s přijetím geostatistiky. Pravděpodobnostní model byl těžko pochopitelný pro ty, kterým byl v době svého zrodu určen, pro geology, prospektory surovin a důlní inženýry, neboť

jevy, kterých se týkal, měly evidentně deterministickou povahu (např. obsah rudy v ložisku) a jejich popis pomocí aparátu náhodných funkcí se jim zdál přespříliš abstraktní. Matheronova práce nebyla zpočátku příznivě přijímána ani statistickou komunitou. Částečně to snad bylo způsobeno vědeckou rivalitou, možná i soupeřením frankofonního a anglosaského světa. Také Matheronovi vytýkali, že obdoby metod, které vyvíjel, jsou již známy v jiných disciplínách, např. v teorii časových řad. A jak už to ve vědě bývá, paralelně s Matheronem, či dokonce o něco dříve, pracovali na podobných myšlenkách další. Švéd Bertil Matérn vyvinul podobnou teorii pro aplikace v lesnictví a Rus Lev Gandin pro meteorologii.

Jelikož Matheron publikoval převážně francouzsky, hlavně v interních zprávách Centre de Géostatistique, a většina jeho žáků byli Francouzi, nepronikala geostatistika příliš rychle do anglosaských zemí, zejména do USA. O rozšíření geostatistiky v Americe se teprve později významně zasloužili dva Matheronovi studenti, André Journel (Stanford University) a Michel David (École Polytechnique de Montreal).

Geostatistika byla od počátku spojena s nástupem počítačů a její rozvoj reagoval na změny v této oblasti. Postupně vznikly komerční balíky programů, například BLUEPACK (1970) a ISATIS (1990) pracující na sálových počítačích a pracovních stanicích. Tyto programy byly drahé, neboť byly primárně určeny pro bohaté těžební společnosti. Volně dostupné programy byly těžko k mání. První volně dostupný balík programů Geo-EAS (1988) byl financován americkou Environmental Protection Agency a posléze uvolněn pro public domain. Existoval již ve verzi pro PC. V roce 1992 André Journel a Clayton Deutsch publikovali knihovnu fortranských programů pro geostatistiku GSLIB. Současné verze těchto programů jsou dostupné na webových stránkách Stanfordské univerzity. Dnes nalezneme programy pro statistiku na běžných osobních počítačích v rámci komerčních programů, například Surfer a ArcGIS. V těchto programech je geostatistika zařazena mezi interpolační metody. Tomuto pojetí odpovídá i tento text. Vedle komerčních programů existuje řada volných programů dosažitelných na Internetu, určených zejména pro prostředí R a Matlab (viz například gstat, geoR, www.ai-geostats.org, cran.r-project.org/web/views/Spatial.html, mgstat.sourceforge.net).

O geostatistice existuje rozsáhlá knižní literatura a články o ní či s jejími aplikacemi se objevují v nejrůznějších odborných časopisech. Teoretičtější práce a články popisující software lze nalézt v časopisech Mathematical Geosciences (dříve Mathematical Geology) a Computers and Geosciences. Oba časopisy vydává vědecká společnost International Association for Mathematical Geosciences, jejíž vznik byl ovlivněn geostatistikou. Pro nás, pro Čechy, není bez zajímavosti, že tato renomovaná vědecká organizace, jejímž prvním prezidentem se stal George Matheron, byla založena na 23. mezinárodním geologickém kongresu konaném v srpnu 1968 v Praze. Geostatistika našla brzy své propagátory i v naší republice, jmenujme alespoň dr. Václava Němce a prof. Ctirada Schejbala.

Matheronova metoda měla původně sloužit k vyhodnocování ložisek nerostných surovin. S jejím rozvojem se však postupně objevily aplikace prakticky ve

všech oborech, které pracují s prostorovými daty. Bez nároků na úplnost můžeme uvést například demografii, ekologii, epidemiologii, geofyziku, geografii, hydrogeologii, kartografii, klimatologii, lesnictví, meteorologii, rybářství a zemědělství (řazeno dle abecedy, nikoliv počtu aplikací). Geostatistika dnes patří mezi základní metody studia prostorových jevů.

Jak již bylo řečeno, záměrem tohoto textu je představit geostatistiku v kontextu prostorových interpolačních metod. Následující výklad se proto dříve, nežli dojdeme k vlastní geostatistice, zabývá různými metodami, které s geostatistikou nějak souvisí.

PROSTOROVÁ INTERPOLACE

V oborech pracujících s prostorovými veličinami se setkáváme s úlohou, kdy máme data změřená v nepravidelné síti bodů (tomu se v angličtině říká *scattered data*) a zajímá nás průběh veličiny mezi nimi, v celé oblasti. Datová síť většinou není dostatečně hustá, dat je málo, a je nemožné zjistit průběh prostorové veličiny přesně. Spíše nežli detaily může být proto rozumné hledat nějakou reprezentativní plochu, jakýsi prostorový průměr závislý na souřadnicích.

Zpravidla se pomocí nějaké interpolační metody dopočtou hodnoty v hustší pravidelné pravoúhlé síti, které se pak buď přímo zobrazí v rastrové grafice či prostorové projekci nebo se jimi proloží izolinie. Interpolačních metod, které k tomu lze použít, je nepřeberně. Jsou založeny na prokládání vhodných funkcí nebo na vážených průměrech. V programech jako Surfer a ArcGIS najdeme například tyto metody:

- *metoda nejbližšího nebo přirozeného souseda,*
- *po částech lineární interpolace,*
- *prokládání (globálních a lokálních) polynomů,*
- *metoda inverzních vzdáleností,*
- *metoda radiálních bazových funkcí,*
- *různé varianty splajnů,*
- *kriging.*

Volba interpolační metody záleží na řadě faktorů, počínaje tím, jaké metody jsou dostupné v tom či onom počítačovém programu a co o nich uživatel ví. Jistě se shodneme, že výběr metody by měl odrážet vlastnosti prostorové veličiny, ať už jsou vyjádřeny prostřednictvím dat anebo dány našimi apriorními představami a teoretickými znalostmi.

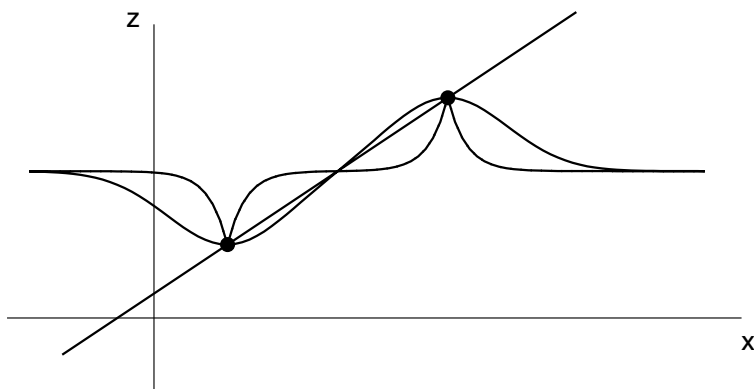
Prostorové veličiny mohou vykazovat velmi rozdílnou proměnlivost. Na otázku, jak vypovídá hodnota naměřená v jistém bodě o hodnotách, které můžeme očekávat v jeho okolí, odpovíme jinak, je-li touto veličinou obsah zlata nebo intenzita gravitačního pole. Zatímco obsah zlata bývá značně proměnlivý až eratický, má intenzita gravitačního pole daleko hladší průběh, neboť se v ní projevují

integrálně hustotní anomálie nacházející se v různých hloubkách. Důležitou roli při studiu prostorových veličin proto hraje rozložení měřených bodů, jejich vzájemná vzdálenost a konfigurace (*geometrie vzorkování*). Uvedený příklad také jistě čtenáře upozornil na roli měřítka, resp. rozměru zpracovávané oblasti.

Volba interpolační metody by měla odrážet naše předpoklady o chování prostorové veličiny. Na obr. 1.1 je jednoduchý příklad. Ve dvou bodech na ose x jsme změřili hodnotu z a proložili tři různé interpolační křivky. Přímka vyjadřuje stoupající tendenci (trend), který pozorujeme v naměřených hodnotách (s rostoucím x roste i hodnota z). Pokud vykreslenou část přímky prodlužujeme nalevo a napravo od bodů měření, vyjadřujeme tím svůj předpoklad, že tato tendence bude pokračovat. Oproti tomu ale můžeme zaujmout stanovisko, že ve větší vzdálenosti od bodů měření nemáme dosti informací, a proto zde za reprezentativní hodnotu raději vezmeme průměr. Takto se chovají obě zbývající křivky. Většine se asi na první pohled zalíbí více ta hladší. Důvodem k tomu může být třeba estetické cítění nebo představa, že průběh dané veličiny nemá mít zuby nebo špičky, ať už v bodech měření nebo mimo ně. Můžeme také klást formální, matematické požadavky na interpolační křivku. Chceme například, aby měla všude derivaci, což je důležité pro další využití v matematických modelech. Křivka se špičkami toto nespĺňuje. Také klesá k průměru rychleji nežli ta hladší, a to dokonce i mezi měřenými body, což se může zdát podivné či zbytečné. Co když je ale měřená veličina tak proměnlivá, že i v blízkosti měřených bodů může nabývat dosti rozdílných hodnot? Potom je přirozené omezit vliv naměřených hodnot jen na nejbližší okolí a netrvat na hladkém průběhu. Na tomto příkladu vidíme, že při volbě interpolace se objevují různá, někdy i protichůdná hlediska:

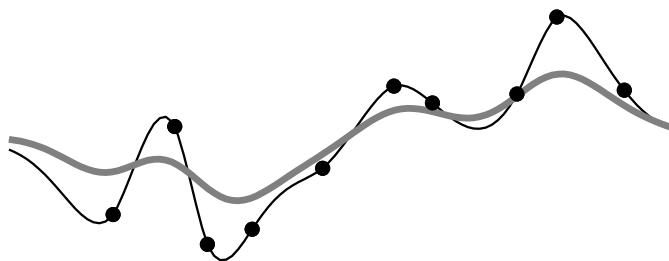
- 1) *průběh interpolace by měl odpovídat naší představě o chování veličiny,*
- 2) *výsledná funkce má mít nějaké formální (matematické) vlastnosti,*
- 3) *chceme minimalizovat chybu, které se při interpolaci dopouštíme.*

Zde můžeme předeslat, že geostatistická metoda kriging byla vytvořena tak, aby především minimalizovala chybu odhadu.



Obr. 1.1: Která interpolace je vhodnější?

Další otázkou je, zda musí interpolace procházet datovými body. Takovým postupům se říká *přesná interpolace* (*exact interpolation*). Přesná interpolace nemusí být vždy vyhovující. Například mohou být data zatížena nezanedbatelnou chybou a je nesmyslné trvat na tom, aby výsledek interpolace v bodě měření byl totožný s naměřenou hodnotou. Nebo při určité konfiguraci bodů měření dává interpolační metoda nerealistické výsledky (záškuby apod.), které jsou spíše odrazem vlastností metody nežli možnou realitou (říkáme jim *artefakty*). Většina interpolačních metod proto nějakým způsobem umožňuje, aby výsledek interpolace (křivka, plocha) datovými body neprocházel. Tento případ nazveme *aproximace*. Oba případy jsou znázorněny na obr. 1.2.



Obr. 1.2: Přesná interpolace (tenká černá) a aproximace (silná šedivá)

Interpolační křivky na obr. 1.1 a 1.2 pokračují ven z intervalu, v němž se nacházejí datové body, doleva i doprava. Striktně vzato, měli bychom v těchto okrajových částech hovořit spíše o *extrapolaci*. Vzhledem k tomu, že se u prostorových dat většinou jedná jen o malý přesah, budeme v následujícím textu hovořit pouze o interpolaci, vědomi si toho, že zpracováváme i tyto okrajové části.

Na tomto místě je vhodné učinit následující poznámku. Jak si čtenář asi povšiml, význam pojmů bod měření, měřený bod a datový bod se může lišit podle kontextu. Abychom předešli nedorozumění v dalším výkladu, pokusíme se o vysvětlení pro případ plošných dat (dat naměřených v ploše, tj. ve 2D), který bude v centru našeho zájmu. Bodem měření máme na mysli místo určené souřadnicemi x a y . Bod v rovině xy určený těmito souřadnicemi můžeme označit jako datový bod. Když ale zobrazíme průběh měřené veličiny jako plochu v prostoru, hodí se označení datový bod pro bod se souřadnicemi (x, y, z) , kde z je naměřená hodnota. To je bod, který leží na interpolační ploše (v případě přesné interpolace) nebo někde nad či pod ní (v případě aproximace). Ať už bude jednat o první nebo druhou variantu, vzájemná vzdálenost datových bodů bude vždy znamenat vzdálenost měřenou v rovině xy , spočtenou ze souřadnic x a y .

Interpolační metody se někdy rozdělují na *deterministické* a *statistické*. Do deterministických metod se zařazují metody založené na formálních požadavcích na interpolační funkci, na její jednoduchost, snadný výpočet, tvar, spojitost, spojitost derivací, vizuální hladkost, nebo na nějakém úsudku, například,