Posudek na diplomovou práci Bc. Ondřeje Krýzy "Aplikace multivariantní statistiky pro analýzu 2D termo-mechanických numerických modelů diapirizmu"

Obsah a členění práce

Diplomová práce Bc. Ondřeje Krýzy "Aplikace multivariantní statistiky pro analýzu 2D termo-mechanických numerických modelů diapirizmu" představuje komplexní pojednání zahrnující vedle geologické látky také matematické, fyzikální, geofyzikální, počítačové a statistické aspekty. Podstatnou roli hraje numerické modelování.

V úvodu práce autor vysvětluje, v čem spočívá význam numerických simulací pro tříbení názorů na regionální geodynamický vývoj litosféry a potřebu efektivního vyhodnocení role velkého počtu parametrů. Uvádí, že "Hlavním cílem práce je porozumět principům a aplikovatelnosti metod multivariantní statistiky při analýze termo-mechanických numerických modelů...a hledání způsobu, kterým lze porovnat jednotlivé simulace...". Pod termínem "multivariantní statistika" autor zřejmě má na mysli vícerozměrné statistické metody (viz poznámka níže).

V následujících kapitolách pak postupně systematicky vykládá potřebné teoretické elementy a dosažené výsledky. Nejprve se zabývá výkladem aparátu parciálních diferenciálních rovnic ve vztahu k mechanice kontinua a termálnímu a mechanickému modelování, včetně popisu metody konečných prvků. Dále popisuje programový systém Elmer v jeho využití pro geofyzikální modelování. Celkem tyto kapitoly, druhá až čtvrtá, mají cca 35 stran.

Geologické a geofyzikální znalosti o českém masivu a moldanubické části jsou shrnuty na 17 stranách páté kapitoly a vytvářejí rámec, do kterého jsou zasazeny numerické exhumační modely a studie středněkorového diapirizmu popisované v nejrozsáhlejší kapitole šesté, která má cca 110 stran. Následuje stručný závěr s poznámkami a naznačením dalších perspektiv.

Vlastní výsledky popsané v práci

Kapitoly 2. – 5. mají charakter spíše sumarizační a prošel jsem je pouze zběžně. Nicméně konstatuji, že jsou podány kultivovaným způsobem a svědčí o tom, že si autor doplnil nadstandardním způsobem znalosti v oblasti matematicko-fyzikálního popisu geodynamických procesů a jejich numerického modelování. Je zřejmé, že je schopen praktické realizace modelů (Elmer) a jejich začlenění do současného tříbení názorů na vývoj Českého masívu (5. kapitola).

Jádro práce a vlastního přínosu autora představuje kapitola šestá. Autor se zde pokouší o statistické vyhodnocení role parametrů ve dvou modelech diapirizmu označených model A a B. Celkem využívá 250 simulací vybraných z většího počtu. Uvádí fundovaný popis simulací včetně

popisu parametrů, které jsou rozděleny na vstupní a výstupní a ty dále na lokální a globální. Vztahy mezi parametry se snaží kvantifikovat pomocí korelační analýzy a metody hlavních komponent. Zmíněna, ale nevyužita, je také kanonická korelační analýza.

Samotná volba parametrů není jednoduchá, lze zde ocenit promyšlený přístup autora. Je zřejmé, že výstupní parametry by v důsledku toho, jak jsou zavedeny, měly vykazovat vzájemnou závislost, nikoliv nutně lineární či jednoduchou. Jelikož vstupní parametry se volí a vztah mezi nimi a výstupními parametry může být složitý, očekával bych spíše v této části nějakou aplikaci vícerozměrné regresní analýzy, nežli přímočaré zjednodušení na charakterizaci vztahu dvojic parametrů prostřednictvím korelačního koeficientu.

Pro studium vztahů mezi parametry autor používá korelační diagram, kde jsou znázorněny bodové grafy a zároveň absolutní hodnota korelačního koeficientu vyjádřená v barevné škále a dále rozptylové diagramy, kde body znázorňující vzájemný rozptyl dvou parametrů jsou barevně škálovány třetím parametrem, popř. velikost bodů čtvrtým parametrem. Toto promyšlené grafické znázornění, které hraje důležitou roli při interpretaci, lze pochválit.

Autor vysvětluje principy korelace i analýzy hlavních komponent. Vzhledem k tomu, že aplikace vícerozměrných statistických metod je hlavním cílem práce, se jeví tento popis nepřiměřeně stručný. Jde pouze o několik stran (4 stránky věnované korelaci a regresi a 1-2 PCA v porovnání s podrobným popisem matematicko-fyzikálních elementů numerických modelů v 2. a 3. kapitole) . Nelze ani říci, že by uvedený popis byl dostatečně výstižný a bezchybný. Při výkladu Pearsonova korelačního koeficientu je zcela opominuta pravděpodobnostní stránka, zejména předpoklady nutné pro korektní použití koeficientu a pro testování závislosti. Autor se tak staví na úroveň heuristického použití korelačního koeficientu jako kvatifikátoru závislosti a volí ad hoc hodnotu 0.5 (nikterak nezdůvodněnou), nad níž považuje korelaci za "významnou". Tato metodická nedostatečnost, nikoliv neobvyklá v geovědách, se mu ale vymstí neboť vzápětí interpretuje, resp. pro interpretaci modelů využívá velikost korelačního koeficientu i v případech, kdy bodové grafy a histogramy vykazují nelinearitu, nenormalitu, odlehlé hodnoty a zejména bimodalitu.

Například v modelu A vidíme v bodovém grafu parametrů Sdom a Tp dva výrazné shluky oba vykazující silnou kvazilineární klesající závislost. Ty "dohromady" dávají $|R| = 0.22$, což je podle autorovy klasifikace nevýznamná hodnota (<0.5) . Další příklad: korelace dP a MP1 je dle autora významná, kdežto dP a MT1 či dT, které vzkazují podobnou závislost dvou shluků (ve skutečnosti v obou shlucích dokonce silnější) významné nejsou. Zdálo by se, že problémy pominou alespoň tam, kde absolutní hodnota korelačního koeficientu je dosti velká (dejme tomu je větší nežli 0.8) a že to svědčí o silném spřažení parametrů. Tak tomu je skutečně všude v matici bodových grafů modelu A, obr. 43. Tuto slibnou myšlenku ale anulují případy shluků v modelu B (např. MT2 a tchar, $|R|\sim 0.9$).

Autor nikde nekomentuje tyto problémy (alespoň jsem na to nenarazil). Jeho použití korelačního koeficientu je metodicky nesprávné.

Při takovémto stavu věcí je zbytečné se podrobněji zabývat výsledky metody hlavních komponent, neboť nedostatečné uchopení dat na úrovni korelace se přímo promítá do této metody. PCA je vystavěna na kovarianční a v případě standardizace (autoškálování, jak to nazývá autor) přímo na korelační matici. Tím pádem přebírá veškeré problémy, na které jsem poukázal při hodnocení korelační analýzy. Nicméně konstatuji, že autor si osvojil metodu PCA.

Není mi z textu jasné, do jaké míry si autor uvedené problémy uvědomuje a myslím, že je to hlavní námět na diskuzi při obhajobě.

V kontrastu s tím, co jsem právě uvedl, ukazuje zasvěcený komentář a rozbor výsledků simulací autorovu schopnost data analyzovat. Domnívám se, že hlavní zdrojem informací pro autora jsou nakonec bodové, resp. rozptylové grafy a určitá jejich syntéza založená na dobré znalosti modelů probíhající v autorově hlavě. Tím jsou překlenuty metodické problémy při zacházení s korelacemi a shrnutí výsledků v paragrafu 6.4 má hlavu a patu.

Styl práce

Celkově je práce podána přehledným a kultivovaným způsobem. V textu se objevují drobné prohřešky proti českému jazyku, které jsou ale z hlediska zaměření práce nepodstatné a podrobněji se jimi proto nezabývám.

Při popisu použitých matematických a jiných nástrojů, které jsou běžně známé, by autor měl dbát na jeho vyváženost. Stačí je přiměřeně připomenout a uvést patřičné citace. Pokud chce zabíhat do detailů, například proto, aby na některém místě zdůraznil stupeň svého porozumění, musí to udělat korektně a dostatečnou mírou vysvětlení, ne jenom něco nadhodit a jít od toho pryč (viz například text pod vzorcem (6.21) nebo vzorec (6.25)) .

Další poznámky

Pojem multivariantní statistika je nezvyklý. Běžně a dlouhodobě se používají přívlastky vícerozměrná nebo mnohorozměrná, popř. bych snesl multivariační. V angličtině se používá multivariate statistics, nikoliv multivariante (to je ve španělštině). Divím se, že se autor nepodíval alespoň na Internet, kde mi vyhledavač google sdělil: vícerozměrná statistika 3 240x, mnohorozměrná statistika 150x, multivariační statistika 51x, multivariantní statistika 5x (z toho 1x Krýza). Přitom autor cituje českou literaturu užívající v názvu termín vícerozměrná data a sám ho používá.

Autor volně zachází s pojmem korelace a občas ho chápe jako závislost. To je běžné, ale je třeba se vyvarovat výroků typu "pro hodnotu (korelačního koeficientu) 0 data nevykazují žádnou korelaci".

Místy zbytečně používá nekonvenční matematický zápis, což je spíše matoucí. Například "Xi(k) je průměr hodnot i(k)-tého vektoru", kde zcela stačí vynechat všude (k).

Přímé zavedení koeficientu determinace v jedné větě s oborem hodnot korelačního koeficientu je nevhodné. Mimochodem, není zde ani vysvětleno, co znamená zápis "D_R...".

PCA se v práci zavádí jen s kovarianční maticí. Přitom se již dříve věnovala pozornost škálování (6.2.2) a na str. 150 se uvádí, že data byla standardizována před vstupem do PCA (resp. autoškálována).

Dodatek, str. 178 – je třeba upřesnit, kdo sestavil skripty. U některých je uvedeno, že jsou sestaveny s pomocí školitele, někde ale nic a není jasné, zda autorem je autor diplomové práce.

- str. vii-viii – vše je anglicky, jen “klíčová slova” česky
- často používá slovo “rozmezí”, většinou ve smyslu interval
- “inciální” – proč ne počáteční?, často jsou tím míněny vstupní parametry
- u termínů, které jsou standardně používány v češtině působí použití česko-anglických (nezaběhaných) termínů křečovitě
- zbytečně píše dvojtečku před každý vzorec
- antikorelace – spíše záporná korelace
- před nebo se nepíše čárka; "nebot'" se nepíše na začátku věty
- nedostatečná velikost fontů v některých grafech, resp. diagramech, např. obr. 46 a následující (kdybych neměl možnost si to v počítači zvětšit, nepřečetl bych je)

Celkové hodnocení

Celkově, přes veškeré kritické výhrady, pokládám předloženou diplomovou práci za velmi kvalitní. Rozhodně může posloužit jako materiál k obhajobě, kde autor zaujme stanovisko k uvedeným problémům. V případě obhájení bych doporučoval, aby v práci byly před jejím definitivním zveřejněním odstraněny formulační nedokonalosti při výkladu statistických metod. Také metodické problémy, na které narazil autor při použití korelační analýzy, by měly být na vhodném místě (v závěru?) uvedeny. Netroufám si na tomto místě přímo navrhnout známku, ta vyplyne až průběhu obhajoby. Můj předpoklad je v první polovině klasifikační stupnice.

Josef Ježek, 1.9.2013