

OPONENTSKÝ POSUDEK NA DIPLOMOVOU PRÁCI

Barbora Lebdušková: Vývoj dynamického modelu pro odhad radonové zátěže budov

Diplomantka se ve své práci zabývá statistickými modely pro analýzu funkcionálních dat. Spíše z názvu než z obsahu diplomové práce soudím, že motivací pro vypsání tohoto tématu byla potřeba sestavit statistický model pro zcela konkrétní netriviální aplikaci, vyšetřit vlastnosti navrženého modelu, případně kriticky porovnat vlastnosti několika navržených modelů a poté tyto modely odhadnout na konkrétních datech a učinit závěry srozumitelné pro osobu, jež data sbírala. Bohužel, práce působí dosti odbitým dojmem s tím, že si z ní nevezme téměř nic ani statistik, kterého zajímá alespoň minimálně teoretické pozadí, ale ani nestatistik, kterého zajímají závěry statistické analýzy na jím sesbíraných datech. „Teoretická“ část je vesměs souhrnem definic a vzorců převzatých z literatury (obvykle bez odvození), na několika místech s chybami či opomenutím uvést předpoklady, za kterých je daný vzorec platný (např. intervaly spolehlivosti). V mnoha případech (např. netriviální optimalizace či cross-validace) mi navíc ani není jasné, jak danou věc spočítat v konkrétní situaci. Domnívám se, že pro tyto případy se autorka plně spolehla na „černou skříňku“ použitého programu bez toho, že by ji zajímalo, jak použitá černá skříňka ve skutečnosti funguje. Též krátká „aplikační“ část je velice vzdálena profesionální analýze (kterou bych očekával u kandidáta titulu Mgr. v matematické statistice) a připomíná spíše špatně strukturovanou ukázkou některých možností Rkového balíčku fda.

Práce začíná jednostránkovou kapitolou nazvanou „Úvod“, z které jsem víceméně vůbec nepochopil, o čem bude práce pojednávat. O datech, která dle mého názoru sloužila jako hlavní motivace pro vypsání tématu DP se mluví v jediné větě: „V závěru jsem model aplikovala na reálná data, která popisují koncentraci radonu v testovacím objektu.“ O několik vět dříve se diplomantka zmiňuje o jakémsi regresním modelu, který umožňuje změnu závislosti (odezvy na regresorech) v čase. Zajímalo by mne tedy, závislost čeho na čem budeme v souvislosti s koncentrací radonu zkoumat a dále důvody, které autorku vedou k tomu, že by se forma této závislosti měla měnit v čase.

Na začátku druhé kapitoly nazvané „Funkcionální data“ dospěla diplomantka k názoru, že je potřeba se zabývat odhadem funkční závislosti odezvy měřené v diskrétních časových okamžicích na spojitém čase. Nutno podotknout, že čtenář v daném okamžiku netuší, proč je toto v kontextu DP užitečné. Diplomantka uvažuje aproximaci funkcionálního pozorování pomocí lineární kombinace vhodných bazických funkcí (v práci je uvažována Fourierova báze a báze založená na B-splinech) a odhad metodou penalizovaných nejmenších čtverců. Bohužel, pro téměř nic z toho, co diplomantka navrhuje použít, se čtenář nedozví, proč by toho či onoho měl použít. V lepším případě je odkázáno na literaturu, v horším případě si musí čtenář sám domýšlet (např. jak minimalizovat výraz (2.10) vzhledem k λ).

Třetí kapitola je nazvána „Analýza funkcionálních dat“ a autorka v ní *uvádí* (nikoliv odvozuje), často s chybami či podstatnými nepřesnostmi, intervaly spolehlivosti pro parametry uvažovaných modelů. Dále (v oddíle 3.2) se pokračuje *popisem* hlavního modelu, kterým by měla být analyzována

motivační data. „Teoretická“ část končí na str. 25 chybným výrazem (3.31) pro interval spolehlivosti regresní funkce $\beta(t)$.

Vrcholem snažení dle mého názoru mohl být oddíl 3.3 (Aplikace na reálná data), na kterém mohla diplomantka alespoň předvést schopnost profesionálně provedené statistické analýzy. Bohužel, oddíl 3.3 se v předložené práci smršknul na cca 2,5 stran textu, jednu tabulku a několik obrázků (jejichž smysl mi v některých případech uniká), přičemž zpráva o analýze vykazuje typické rysy pro zprávu vytvořenou nestatistikem, který pouze „mačká“ tlačítka statistického programu, data přizpůsobuje dostupné softwarové implementaci (nikoliv naopak) a v konečném důsledku je pro něj důležité pouze množství zaplněného papíru, kde se obsah podané informace řídí zejména obsahem přímého výstupu z použitého programu a nikoliv cílem analýzy. Na tomto místě DP jsem se sice dozvěděl, že budeme modelovat závislost objemové aktivity radonu v dětském pokoji na objemové aktivitě radonu v chodbě, nicméně zůstalo mi utajeno, k čemu má takový model sloužit a jaké jsou závěry z provedené analýzy.

Čtvrtá kapitola, která popisuje použitý Rkový balíček fda Ramsaye, Wickhama, Gravesa a Hokera by dle mého názoru měla být uvedena pouze ve formě appendixu. Diplomová práce končí půlstránkovou pátou kapitolou nazvanou „Shrnutí“, z něhož již první věta („*Podářilo se nám teoreticky popsat metody. . .*“) není dle mého názoru pravdivým výrokiem.

Byť mi není znám žádný striktní limit pro minimální rozsah diplomové práce, domnívám se, že s ohledem na tristní obsah, je předložená diplomová práce nedostatečná též co do svého rozsahu. Práce je vysázena v textovém okně, s kterým by ji bylo možné tisknout na formát A5. Odečteme-li titulní stránku, obsah, abstrakt a seznam literatury, vešla se diplomová práce na 37 stran. Z těchto 37 stran je však 5 stran celoplošně pokryto obrázky a 6 stran zabírá popis použitého software (volně dostupný Rkový balíček fda obšírně zdokumentovaný též v knize napsané podmnžinou autorů balíčku a vydané nakladatelstvím Springer). Další obrázky a tabulky nalezneme též na zbylých 26 stranách.

V předložené podobě práci bohužel dle mého názoru nelze uznat jako práci diplomovou pro studijní směr matematická statistika na Matematicko-fyzikální fakultě Univerzity Karlovy v Praze.

RNDr. Arnošt Komárek, Ph.D.
oponent diplomové práce

Výběr konkrétních připomínek a dotazů

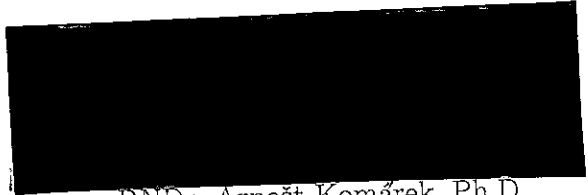
- str. 6 dole Nebylo by vhodné předpokládat něco o střední hodnotě $\epsilon(t_i)$?
- str. 7 dole „V tomto případě je dobré použít co největší bázi.“ Jak velkou?
- str. 8 nahoře Co se myslí „*vysokým rozptylem křivky*“?
- str. 8 Tvrdí se zde, že optimální λ minimalizuje výraz (2.9), respektive (2.10). Na další straně se hovoří o „*výše uvedené automatické optimalizaci parametru λ* .“ Zajímalo by mne, jakým způsobem lze výraz (2.9), respektive (2.10) minimalizovat vzhledem k λ v konkrétní situaci.
- str. 8 Jak souvisí GCV (výraz (2.10)) s cross-validací? Na první pohled nijak, proto bych uvítal vysvětlení.
- str. 14 Jak jsou voleny uzly pro spliny na obr. 2.4?
- str. 16, Tabulka 2.1 Domnívám se, že tabulka uvádí hodnoty MMSE a nikoliv MSE. Dále by mě zajímalo, jak jste spočítala hodnoty zde uvedené, tj., jak jste spočetla integrál ve vzorci (2.16)?
- str. 19 „V článkách [13] a [8] je dokázáno, že splajny jsou Bayesovské odhady s Gaussovou apriorní hustotou, ...“ Pro co se uvažuje ona Gaussova apriorní hustota? V jakém smyslu Bayesovské odhady?
- str. 19 uprostřed Z čeho plyne, že $\hat{y}(t_i) \pm z_{\alpha/2} \hat{\sigma} \sqrt{(S_{\lambda\Phi})_{ii}}$ je intervalem spolehlivosti pro $y(t_i)$?
- str. 19 dole Nemyslím, že by $\sigma_e^2(t_j)$ bylo rozptylem odhadů $\hat{y}(t)$.
- str. 23 „... ale obecně je vhodné tuto soustavu řešit numerickými metodami integrace.“ Jakými? Jak?
- str. 25 výraz (3.31) Nevím, co znamená index ii u matice Θ . Navíc, i kdybych nahradil ve výrazu (3.31) matici Θ_{ii} maticí $\Theta(t)$, obávám se, že stále se nebude jednat o správný tvar, jestliže víme, že $\hat{\beta}(t) = \Theta(t)\hat{b}$. Konečně by mě také zajímalo, proč by zrovna toto (s kvantilem normovaného normálního rozdělení) měl být interval spolehlivosti s patřičnou pravděpodobností pokrytí.
- str. 29 dole „Z údajů v tabulce 3.3 je vidět, že...“ Bohužel, co bych dle dalšího textu v tabulce měl vidět, opravdu nevidím.
- str. 30 nahoře „Pro první a poslední den není k dispozici dostatek dat, proto je z další analýzy vyřadíme.“ Nebylo by vhodnější odhadnout pro první a poslední den pouze část funkcionální křivky a tu použít v „concurrent“ modelu?
- str. 30 uprostřed „To je při výpočtech velmi nepraktické, proto chybějící hodnotu nahradíme lineární aproximací.“ Osobně několik dodatečných řádek ve skriptu pro analýzu nepovažuji za nepraktičnost.

- str. 30 dole „*Problém je v tom, že hodnoty nebyly měřeny pro všechny dny ve stejných časech...*“
Zajímalo by mě, v čem spočívá tento problém.
- str. 31, Obrázek bez popisky Lepší (a klasičtější) znázornění funkcionálních dat je dle mého názoru ve formě po částech lineárních funkcí (jedna pro každý den).
- str. 31 Testujete normalitu na reziduích. Proč to děláte, když o nějaké normalitě není v celé „teoretické“ části práce ani zmínka?
- str. 31 W test normality je obvykle nazýván dle svých tvůrců (Shapiro a Wilk).
- str. 32 Co si mám odnést z levé a prostřední části obrázků 3.7 a 3.8?
- str. 33 dole Co myslíte „*podrobnější analýzou rozptylu*“?
- str. 34, Obrázek 3.10 Která čára je M1 a která M2?
- str. 41 „*Struktura modelu byla úmyslně volena jednoduše, aby výsledky byly jasně interpretovatelné a ověřitelné.*“ Jak ověřitelné?
- str. 42 Liší se nějak reference [5] od reference [10]?

NÁVRH HODNOCENÍ DIPLOMOVÉ PRÁCE

Barbora Lebdušková: Vývoj dynamického modelu pro odhad radonové zátěže budov

Navrhuji hodnocení nevyhověla.



RNDr. Arnošt Komárek, Ph.D.
oponent diplomové práce

