

Posudek bakalářské práce

Michal Polák – Zpracování dat z oblasti energetiky pomocí shlukové analýzy

Předloženou bakalářskou práci hodnotím jako zajímavou zprávu o zajímavé analýze dat. Michal se při ní naučil a použil několik matematicko-statistických metod a postupů:

- Hierarchické a nehierarchické metody shlukové analýzy a k tomu potřebné vyjádření vzdáleností dvou funkcí (pomocí různých metrik). Rád bych poznamenal, že jsem už četl několik bakalářských prací týkajících se (alespoň částečně) shlukové analýzy. V Michalově práci jsem konečně na konkrétním příkladě viděl dramaticky rozdílné chování různých shlukovacích metod. Překvapil mě relativně malý vliv zvolené metriky.
- Stanovení konfidenčních intervalů, a to jak parametricky (na základě centrální limitní věty), tak i neparametricky (pomocí odhadu kvantilů). Rád bych vyzdvihl Michalovu důslednost, když našel mně neznámé vylepšení konstrukce intervalového odhadu na základě centrální limitní věty (viz strana 25 dole) uvedené v knize Anděl, J.: Základy matematické statistiky, Praha: Matfyzpress, 2007. (V preprintu této knihy z roku 2002 toto vylepšení zmíněno není).

K dosaženým výsledkům bych řekl, že je jich v práci opravdu hodně – některé jsou zajímavé, jiné (dle mého názoru) méně. Pro mě zajímavý, v textu poněkud zapadlý, výsledek je například uveden na straně 18. Ukazuje, že v prvních sedmi hodinách je úroveň systémové odchylky v podstatě stejná u dnů „průměrných“, kdy její hodnoty po celý den kolísají kolem nuly, jako u dnů „podprůměrných“, kdy převládají záporné hodnoty systémové odchylky. Tento výsledek ukazuje na obtížnost predikce denního průběhu v prvních hodinách dne. Je z něj rovněž patrná osmá hodina ranní coby čas, v němž se denní průběhy začínají výrazně lišit. To by ostatně moho být potvrzeno výpočtem směrodatných odchylek zjištěných hodnot v jednotlivých denních hodinách.

Výsledky práce hodnotím možná poněkud subjektivně jako její „zadavatel“, který si je vědom toho, že zajímavost výsledků je do značné míry dána použitou metodou. Zde jsme chtěli zkusit něco nového. Na začátku práce byla myšlenka, že průběh systémové odchylky ve dvou po sobě jdoucích dnech bude podobný a tudíž, že po „nadprůměrném“ dni bude spíše následovat opět „nadprůměrný“ den a pod. Tuto myšlenku potvrzuje například výsledek na straně 26. V dalších kapitolách se nicméně ukazuje, že možnost využít tuto myšlenku pro predikci budoucího vývoje je malá. Michal to sám v závěru (na straně 43), dle mého názoru trefně, hodnotí slovy „Ukázalo se, že strategie založená na informaci ze shlukové analýzy není efektivní. Jako určující se ukázala informace o znaménku systémové odchylky. Přidání informace o shlucích prakticky moji sázku nijak nemění a vždy tipuju znaménko z předchozí hodiny. Jako užitečná se ukazuje především při hodinovém přístupu, kdy mi pomůže určit situace, kdy se dopouštím největší chyby ve své predikci.“ S trochou nadsázky bychom tedy mohli říci, že převrat v predikci časových řad se nekoná. Tato práce pomohla, přinejmenším mně osobně, objasnit, že předpovídání pomocí shlukové analýzy má jen velmi omezené možnosti.

K textu mám ještě tři drobné připomínky:

- Na straně 21 činí autor rozhodnutí, zda bude nadále pracovat se shlukováním pomocí Wardovy metody či pomocí metody PAM. Uvádí „jako rozumnější volba se jeví Wardova metoda...“. Zdůvodňuje to hodnocením velikostí vzniklých shluků. O tomto rozhodnutí je možné dále diskutovat. Bylo by zajímavé podívat se, jaké výsledky by přineslo použití metody PAM.

- Na straně 22 autor uvádí, že „ve shluku s průměrnými dny jsem žádnou užitečnou informaci nezískal.“ Dle mého konfidenční interval přináší užitečnou informaci i zde (byť třeba ne až tak zajímavou).
- Na straně 23 je poněkud matoucí označení $B \in A$. Symbol A měl být psán jiným fontem.

Práci navrhuji k obhajobě a doporučuji hodnotit stupněm A.

Ondřej Vencálek