České vysoké učení technické v Praze
Fakulta elektrotechnická
Katedra matematiky

Semestrální práce z předmětu X01MVT
Matematika pro výpočetní techniku

Vzdálenost k nejbližší benzinové čerpací stanici χ^2-test dobré shody

Jan Skalický

leden 2008

Obsah

Seznam tabulek

Seznam obrázků

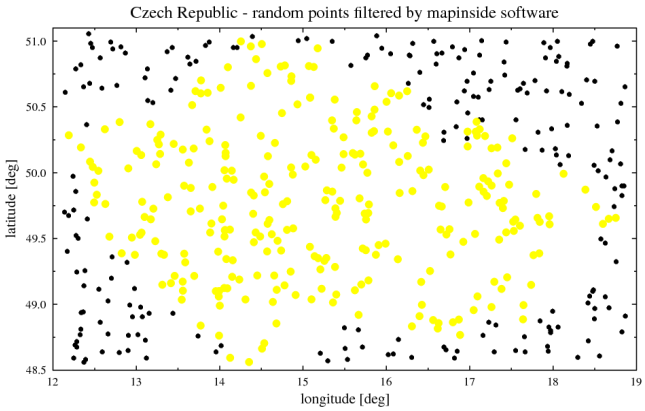
1. Zadání

Otestujte hypotézu, že nejkratší vzdálenost po silnici z náhodně zvoleného bodu v ČR k nejbližší benzinové čerpací stanici má logaritmicko-normální rozdělení. Stanovte jaký minimální dojezdový rádius by si měl řidič nechávat jako rezervu, aby mu i při neznalosti místních poměrů zůstala statistická pravděpodobnost alespoň 90 % na dojetí k čerpadlu pro doplnění paliva. Předpokladem je, že řidič se po mapě pohybuje rovnoměrně.

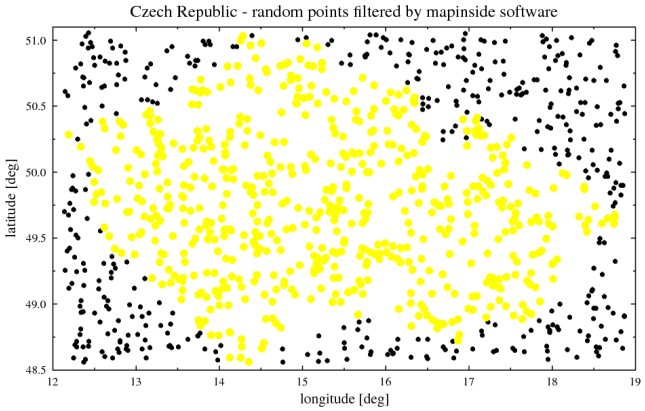
2. Získání dat

Bylo třeba získat data s délkou trasy z náhodně zvoleného bodu na silniční síti ČR k nejbližší čerpací stanici. Za tímto účelen byl vytvořen program, který po předložení obrázku s šablonou mapy území filtruje vstup geografických souřadnic na body ležící uvnitř obrysu. Program provádí lineární korekci souřadnic na sférickém povrchu. Vzhledem k tomu a k rozlišení použitého obrázku mapy (1921*1102 pix) jsme obdrželi přesnost rozhodnutí o hranici lepší než cca 1.5 km, což plně postačuje našemu účelu.

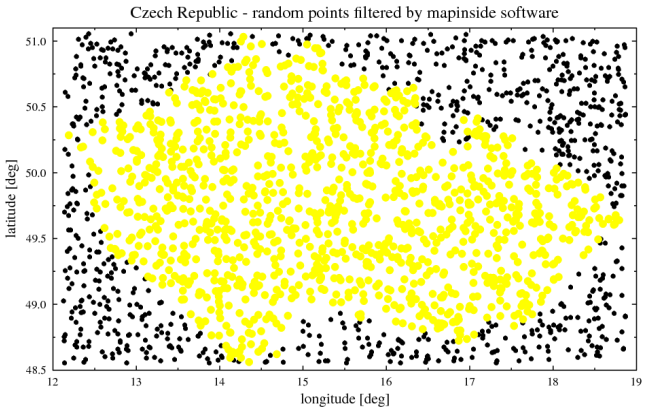
Pomocí generátoru v linuxovém souboru `/dev/random` byly vygenerovány náhodné souřadnice bodů z obdélníka opsaného obrysu mapy ČR a ty následně filtrovány výše zmíněným programem (pro velký počet je to cca 57.7 % bodů z použitého intervalu o rozměrech cca 484*279 km, tozn. pokud bychom takto metodou Monte Carlo integrovali plochu ČR, obdrželi bychom výsledek cca 77.9 tis. km^2). Vizualizaci výsledků této filtrace ukazují obrázky 2.1 až 2.4.



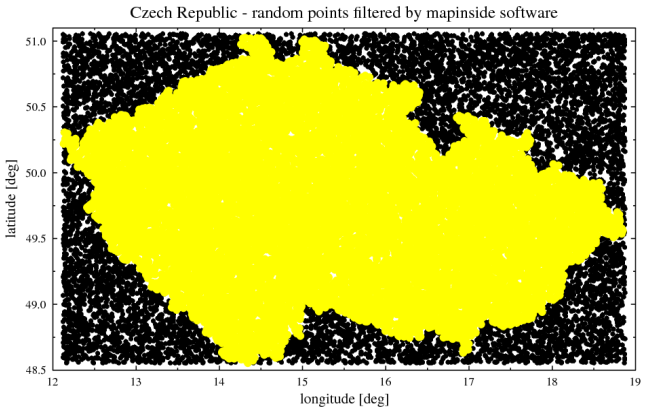
Obrázek 2.1: filtrace pro 500 bodů



Obrázek 2.2: filtrace pro 1000 bodů



Obrázek 2.3: filtrace pro 2000 bodů



Obrázek 2.4: filtrace pro 20000 bodů

2.1 Sběr a vyčištění dat

Pro sběr vzdáleností byl použit soubor 2000 bodů, z nichž 1145 leželo uvnitř ČR. Vlastní vzdálenosti byly získány z veřejného navigačního portálu na www.mapy.cz. Plánovač trasy měl nastaven atribut použití nejkratší cesty a klíčem k vyhledání cílových bodů byl řetězec „benzinová čerpací stanice“. V některých případech byly nalezeny kombinované stanice LPG. Ačkoliv bylo snahou proces sběru co nejvíce automatizovat, bylo nutné provést manuální vyčištění datového souboru, a sice korekce nepřesně nalezených cest

- v důsledku velké vzdálenosti počátků k nejbližšímu bodu na silnici (nejbližší čerpací stanice se vybírá podle přímé vzdálenosti)
- v důsledku jednosměrnosti dálnic a granularity jejich křížení s jinými cestami
- vyřazení počátků na území vojenských újezdů

2.2 Stručné charakteristiky dat

Získaná a vyčištěná data obsahují vzdálenosti od 0.1 km do 33.8 km. Po vyřazení počátků ve vojenských újezdech jsou největší vzdálenosti způsobeny počátky v horách (zejm. Beskydy, Orlické hory) a např. za přehradou Lipno, kterou je nutno celou objet. Vzdáleností přesahujících 20 km je celkem 14.

Z 1145 návštěv bylo 661 stanic různých a 384 z nich bylo navštíveno pouze jednou. 21 stanic bylo navštívenou více než 4x, z toho 2 nejvíckrát navštívené byly 9x a 8x. Histogram zastoupení jednotlivých firem provozujících veřejné čerpací stanice je v tabulce 2.1.

Název firmy	Počet návštěv
Benzina, s.r.o.	224
Čepro, a.s.	214
PAP Oil čerpací stanice, s.r.o.	91
Shell Czech Republic, a.s.	64
Robin Oil, s.r.o.	60
OMV Česká republika, s.r.o.	54
Agip Česká republika, s.r.o.	25
KM - PRONA, a.s.	21
Hunsgas, s.r.o.	16
Svam CS, s.r.o.	14
Inteko Konice, a.s.	13
ostatní (179 subjeků, méně než 10 návštev)	349 (z toho 105 subjektů právě 1x)

Tabulka 2.1: Histogram zastoupení navštívených firem

Podle zběžného porovnání histogramu z tab. 2.1 se zprávou o síti čerpacích stanic PHM v ČR za 1. pololetí 2007, obsahující výsledky statistického zjišťování odboru surovinové a energetické politiky ministerstva průmyslu a obchodu, se zdá, že sesbíraná data jsou v souladu se skutečností.

3. Testovaná hypotéza

Nulovou hypotézu, kterou chceme testovat formulujeme takto: „Nasbíraná data mají logaritmicko-normální rozdělení.“ Alternativní hypotéza toto rozdělení popírá.

Požadujeme test s hladinou významnosti 0.01. To je max. pravděpodobnost chyby 1. druhu – toho, že hypotézu neoprávněně zamítneme, ačkoliv bude platit.

4. Test dobré shody rozdělení

O neplatnosti testované hypotézy se pokusíme rozhodnout provedením χ^2-testu dobré shody. Testujeme na logaritmicko-normální rozdělení, které odpovídá zobrazení náhodné veličiny s normálním rozdělením exponenciální funkcí. Parametry normálního rozdělení můžeme jednoduše odhadnout z realizace výběru, a proto budeme testovat logaritmus původních dat na normální rozdělení. Za účelem snadné modifikovatelnosti a získání přesných výsledků byl celý test naprogramován v matematickém systému Maple 9.5

Z dat (n = 1145), již zlogaritmovaných (základ logaritmu není podstatný, protože má pouze vliv multiplikativní konstanty – na rozptyl výsledného rozdělení a byl použit $\sqrt[10]{10}$, takže nová data dostala význam dBkm), spočítáme výběrový průměr a výběrový rozptyl pro odhad parametrů odpovídajícího normálního rozdělení:

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 7.4889$$

$$\widehat{\sigma^2} = s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2 = 12.385$$

$$\hat{\sigma} = \sqrt{\widehat{\sigma^2}} = 3.5192$$

Test pracuje s diskrétními hodnotami, a tak rozdělíme data do disjunktních tříd s blízkými významy. Všechny teoretické četnosti musí být velké a nejlépe podobné. Ke zvoleným intervalům <a, b) spočítáme teoretické pravděpodobnosti pro testované rozdělení z jeho distribuční funkce a jim odpovídající očekávané četnosti:

$$p_i = F_{N(\hat{\mu},\hat{\sigma}^2)}(b) - F_{N(\hat{\mu},\hat{\sigma}^2)}(a)$$

$$np_i = p_i * n$$

Testovacím kritériem je statistika T:

$$T = \sum_{i=1}^{k} t_i = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

a testujeme ji proti zvolenému (1–0.01) kvantilu χ^2 rozdělení s tolika stupni volnosti, kolik je intervalů diskretizace – 1 (poslední je doplňkem do celku) – 2 (2 neznámé parametry rozdělení jsme odhadovali s použitím stejného souboru dat).

$$q_{\chi((k-1)-2)}(0.99) = 16.812 \; pro \; k = 9; \;\; 9.210 \; pro \; k = 5; \;\; 13.277 \; pro \; k = 7$$

Aktuálně dosažený koeficient spolehlivosti testu lze vyjádřit jako:

$$F_{\chi((k-1)-2)}(T)$$

4.1 Realizace testu

Tabulka 4.1 obsahuje hodnoty testu pro zvolenou diskretizaci dat. Vidíme, že hodnota kritéria je větší než zvolený kvantil (44.588 > 16.812), a proto můžeme na zvolené hladině významnosti 0.01 zamítnout nulovou hypotézu a přijmout hypotézu alternativní.

Po naprogramování testu, včetně dělení do intervalů, nám nečiní potíže vyzkoušet i jinou diskretizaci dat, s důrazem na podobnost teoretických četností, viz tabulku 4.2. Zde rovněž zamítáme (pro 43.646 > 9.210).

i	**Interval [dBkm]**	**n_i**	**p_i**	**np_i**	**t_i**
1	(-∞..3)	109	0.1010614636	115.7153758	0.3897171990
2	<3..5)	112	0.1386536014	158.7583736	13.77152872
3	<5..6)	91	0.0964085430	110.3877817	3.405142067
4	<6..7)	132	0.1086361253	124.3883635	0.4657751624
5	<7..8)	125	0.1129800837	129.3621958	0.1470967007
6	<8..9)	147	0.1084422370	124.1663614	4.199004028
7	<9..10)	144	0.0960647192	109.9941035	10.05471846
8	<10..11)	120	0.0785412865	89.92977304	10.05471846
9	<11..+∞)	165	0.1592119403	182.2976716	1.641323447
∑	9	1145	1.0	1145	**44.58760569**

Tabulka 4.1: Hodnoty testu

i	**Interval [dBkm]**	**n_i**	**p_i**	**np_i**	**t_i**
1	(-∞..4.5)	192	0.1978582449	226.5476904	5.268395850
2	<4.5..6.6)	199	0.2024405794	231.7944634	4.639786534
3	<6.6..8.4)	225	0.2018460258	231.1136995	0.1617269840
4	<8.4..10.5)	318	0.2017525921	231.0067180	32.76022091
5	<10.5..+∞)	211	0.1961025578	224.5374287	0.8161756232
∑	5	1145	1.0	1145	**43.64630590**

Tabulka 4.2: Hodnoty testu pro jinou diskretizaci

Z obou výsledků je patrné, že rozdělení se liší zejm. v oblasti nedaleko za odhadnutou střední hodnotou, a to tak, že hustota klesá rychleji, než podle logaritmicko-normálního rozdělení. Můžeme tedy vyzkoušet ještě test proti normálnímu rozdělení (tozn. původní data nebudeme logaritmovat a bude nutné přepočítat odhady parametrů rozdělení – EX=7.2358 a DX=21.708) v domnění, že záporné výsledky, které data neobsahují, budou mít nízký vliv na příspěvek do testované statistiky. Takovou situaci, pro 7 intervalů dělení zachycuje tabulka 4.3.

4.2 Test na normální rozdělení

V testu shody původních dat na normální rozdělení vidíme opačný trend – přebytek dat pod střední hodnotou a jejich nedostatek v jisté vzdálenosti nad ní. Celkově

zamítáme (porovnávaný kvantil je 13.277), ale k výraznému zhoršení oproti předchozím testům nedošlo. Toto pozorování mě vede k podezření, že naše data mají rozdělení blízké nějaké směsi normálního a logaritmicko-normálního rozdělení. Takové rozdělení by mělo 5 stupňů volnosti a tudíž by bylo vhodné ho testovat při větším počtu intervalů dělení.

i	**Interval [dBkm]**	**n_i**	**p_i**	**np_i**	**t_i**
1	(-∞..2.3)	144	0.1447176118	165.7016655	2.842230246
2	<2.3..4.6)	247	0.1410752921	161.5312095	45.22292734
3	<4.6..6.4)	178	0.1430237912	163.7622409	1.237854240
4	<6.4..8.1)	153	0.1447568750	165.7466219	0.9802695707
5	<8.1..9.9)	133	0.1427009426	163.3925793	5.653309841
6	<9.9..12.2)	116	0.1403900974	160.7466615	12.45602053
7	<12.2..+∞)	174	0.1433353899	164.1190214	0.5948959314
∑	7	1145	1.0	1145	**68.98750770**

Tabulka 4.3: Hodnoty testu původních dat na normální rozdělení

5. Rezerva paliva

Pro stanovení minimálního dojezdu jako rezervy, aby řidiči zůstala statistická pravděpodobnost alespoň 90 % na dojetí k benzinovému čerpadlu, potřebujeme znát rozdělení použité náhodné veličiny. My jsme však ověřovaná rozdělení zamítli, takže musíme vystačit s empirickým rozdělením z realizace výběru. Jeho 0.9-kvantil nám odpovídá hodnotou 13.7 km.

Pro zajímavost můžeme zkusit vyčíslit tento kvantil i z ověřovaných rozdělení. Pro normální rozdělení se zjištěnými parametry činí cca 13.2 km. Pro původní logaritmicko-normální je to 12 dBkm a pro obdržení vzdálenosti musíme provést zpětnou transformaci delogaritmováním a vyjde cca 15.8 km. Je zde opět vidět stejný trend jako v příspěvcích k testovací statistice a i tato statistika nám napovídá, že lépe odpovídající rozdělení bude někde mezi těmito dvěma.

Praktická rada je tedy rezervovat si palivo na dojezd cca **15 km** k dosažení více než 90 % pravděpodobnosti, že budeme schopni ho kdekoliv doplnit.

6. Závěr

Hypotéza, že nejkratší vzdálenost po silnici z náhodně zvoleného bodu v ČR k nejbližší benzinové čerpací stanici má logaritmicko-normální rozdělení, byla na hladině významnosti < 1 % zamítnuta. Rovněž normální rozdělení této veličiny bylo s touto hladinou zamítnuto. Je zde domněnka, že rozdělení by lépe odpovídalo směsi těchto 2. Byla vyslovena praktická rada, nechávat si rezervu paliva na dojetí 15 km.

7. Zdroje

[1] Webová stránka podpory výuky: http://cmp.felk.cvut.cz/~navara/MVT/

[2] Mirko Navara – Matematika pro výpočetní techniku (přednášky předmětu)

[3] Vladimír Rogalewicz – Pravděpodobnost a statistika pro inženýry, 2000