Projekt z manažmentu dát - Analýza horoskopov

Matúš Zubčák

Fakulta matematiky, fyziky a informatiky, UK BA

1 Úvod

V našom projekte sme sa rozhodli analyzovať horoskopy zo stránky `https://diva.aktuality.sk/horoskopy/denny-horoskop/`. Presnejšie, na základe počtu opakujúcich sa horoskopov sme sa rozhodli určiť najpravdepodobnejšiu veľkosť množiny horoskopov na webstránke. Súčasťou nášho projektu je aj porovnanie nami pozbieraných dát s náhodnou vzorkou využitím Monte-Carlo metódy. Následne pre relevantné veľkosti náhodných množín pomocou grafického vykreslenia ukazujeme odchýlku pre túto funkciu náhodnej premennej. Použitým pravdepodobnostným metódam sa neskôr detailne venujeme v našom reporte. Súčasťou reportu je aj popis použitých nástrojov na získanie, uchovanie a následnú analýzu získaných dát.

2 Zber dát

V rámci nášho projektu sme sa rozhodli, že budeme systematickým spôsobom každý deň sťahovať texty denných horoskopov z vyššie uvedenej webstránky, aby sme ich následne na konci semestra mohli porovnať s našimi hypotézami.

Naprogramovali sme v polovici semestra *crawler* podobný tomu z cvičení, ktorý by vedel sťahovať texty horoskopov zo stránky a ukladať ich do našej internej databázy. Pomocou crawlera sme zozbierali dáta od 29. marca do 17. mája, čím sme počas 50 dní vyzbierali 600 textov horoskopov (pre všetky znamenia dokopy). O každodenné automatizované spúšťanie crawlera na serveri tohto predmetu sa stará linuxový príkaz *crontab*.

3 Analýza dát

Zo získaných dát nie je ťažké zistiť, že celé texty horoskopov sa na stránke v čase opakujú. Hodnoty jednotlivých opakovaní je možné si pozrieť v nasledovnej tabuľke:

Počet rôznych horoskopov	Počet opakovaní
429	1
67	2
9	3
1	4

Našou základnou pracovnou hypotézou pri tomto projekte bolo, že autori stránky disponujú databázou horoskopov nejakej fixnej veľkosti, a že z tejto databázy každý deň vyberajú náhodným výberom 12 rôznych horoskopov, ktoré priradia jednotlivým znameniam. Teraz si vysvetlíme, akým spôsobom sme sa rozhodli odhadovať veľkosť množiny všetkých horoskopov na základe získaných dát. V našom projekte predpokladáme, že výber je náhodný a vybrané vzorky sú po dvoch nezávislé.

Vyššie zmienenú úlohu si do matematického zadania môžeme pretransformulovať nasledovným spôsobom: *Ak pri n náhodných po dvoch nezávislých výberoch z množiny U vyberieme m rôznych objektov, aká je potom najpravdepodobnejšia veľkosť $k = |U|$ množiny U?*

Kým prejdeme k riešeniu tejto úlohy, vyriešime súvisiacu podúlohu: *Aká je pravdepodobnosť, že z množiny veľkosti k vyberieme pri n-násobnom náhodnom výbere práve m rôznych objektov?*

Túto úlohu vyriešime ako pomer počtu priaznivých a všetkých možností. Nech $X_{k,\,n}$ je náhodná premenná symbolizujúca počet rôznych objektov pri n-násobnom výbere z k prvkovej množiny s opakovaním, M je počet postupností dĺžky n, v ktorých je práve m rôznych symbolov a $V_n(k)$ sú variácie s opakovaním n-tej triedy k prvkov. Potom:

$$Pr[X_{k,\,n} = m] = \frac{|M|}{V_n(k)} = \frac{\binom{k}{m} \sum_{i=0}^{m-1} F(m,\, n,\, i)}{k^n} = \frac{\binom{k}{m} \sum_{i=0}^{m-1} (-1)^i \cdot \binom{m}{m-i} \cdot (m-i)^n}{k^n}.$$

Pristavme sa teraz pri jednotlivých krokoch. Je zrejmé, že všetkých možností je $V_n(k) = k^n$. Zaujímavejšie je to s výpočtom veľkosti množiny M. Ten vieme vypočítať tak, že najprv si zafixujeme, ktorých konkrétnych m symbolov zo všetkých k vyberáme (k nad m) a následne pomocou princípu inklúzie a exklúzie vyriešime úlohu *koľko existuje postupností dĺžky n takých, že použijeme práve m písmen.* Funkcia $F(m,\, n,\, i) = (-1)^i \cdot \binom{m}{m-i} \cdot (m-i)^n$ vystupujúca vnútri sumy vyjadruje počet postupností dĺžky n z $m-i$ písmen (nemusíme ich použiť všetky).

Keď už vieme túto úlohu vypočítať, potom je ľahké vyriešiť pôvodnú úlohu tak, že maximalizujeme pravdepodonosť $Pr[X_{k,\,n} = m]$ cez k, lebo v podstate len hľadáme takú veľkosť k množiny U, aby výber n objektov, pričom m ich je rôznych, bol čo najpravdepodobnejší. Preto v našom projekte riešime úlohu:

$$\max_k Pr[X_{k,\,n} = m].$$

Môžeme si všimnúť, že suma $\sum_{i=0}^{m-1} (-1)^i \cdot \binom{m}{m-i} \cdot (m-i)^n$ nie je závislá od k, a teda nám stačí riešiť úlohu

$$\max_k \binom{k}{m} / k^n.$$

Túto úlohu riešime algoritmicky. V našom projekte sme si naprogramovali funkciu, ktorá pre zadané m, n vypočíta k, na ktorom výraz nadobúda maximum. Uvedomme si, že ak $m = n$ (všetky výbery boli rôzne objekty), tak

optimálne je dosadiť za $k = \infty$. Inak sa maximum nadobúda vždy pre nejaké konkrétne prirodzené číslo k. To, že k je prirodzené číslo zjednodušuje nájdenie maxima.

Funkcia, ktorú používame postupne vyskúša veľkosti $k = m, \ldots, 10000$ a následne vypíše k, pre ktoré sa nadobúda maximum výrazu. Pre urýchlenie výpočtu sme túto funkciu naprogramovali v programovacom jazyku $c{+}{+}$ a program v pythone, ktorý sa stará aj o vykreslenie výsledných grafov si ju len volá.

Pre lepšiu numerickú stabilitu sme sa rozhodli počítať logaritmus danej hodnoty a nie hodnotu samotnú, maximum ako také to zrejme nijak neovplyvní. Preto v programe počítame

$$\max_k \sum_{i=k-m+1}^{k} \log i - \sum_{i=1}^{m} (\log i) - n \log k.$$

Výsledkom tejto časti výskumu je graf, ktorý ukazuje najpravdepodobnejšiu veľkosť množiny horoskopov v závislosti od nazbieraných dát v čase. Os x predstavuje počet dní od začiatku zbierania dát a os y predstavuje najpravdepodobnejšiu veľkosť množiny horoskopov v závislosti od už nazbieraných dát. Teda napríklad ak $x = 10$, potom sa pýtame aká je najpravdepodobnejšia veľkosť množiny horoskopov na základe prvých 120 nami zozbieraných horoskopov. Graf je možné si pozrieť na obrázku 1.

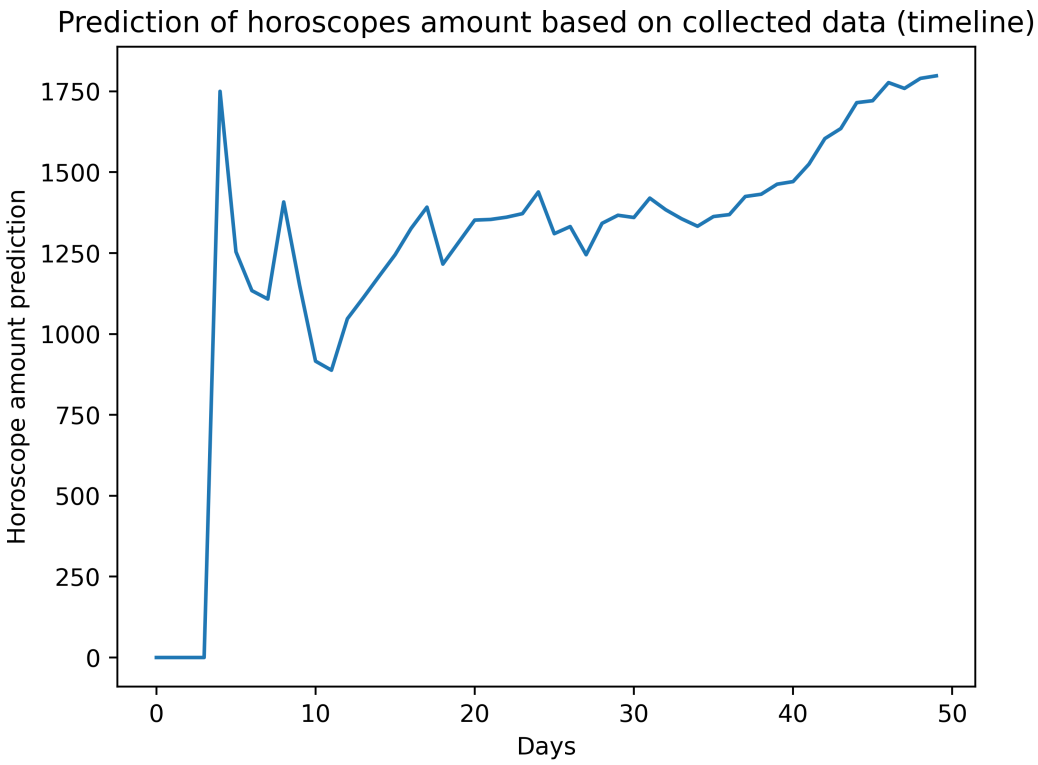
Vieme, že funkcia $\binom{k}{m}/k^n$ môže niekedy nadobúdať maximum pre $k = \infty$. Ak sa tak stane, vykresľujeme to kvôli prehľadnosti v grafe ako hodnotu 0. Tak postupujeme aj vo všetkých nasledujúcich grafoch. Presnejšie, ak nám vo funkcii vyšlo ako najlepšie $k > 3000$, tak zapíšeme 0.

Pre lepšiu interpretáciu výsledkov sme sa rozhodli spraviť ešte Monte-Carlo simuláciu. Cieľom tejto simulácie je pre rôzne veľkosti N množiny horoskopov U odhadnúť priebeh funkcie $\max_k \binom{k}{m}/k^n$.

Pre fixné N v jednom behu simulácie postupujeme tak, že si vygenerujeme postupnosť náhodných čísel veľkosti $0 \ldots N-1$. Následne sa na ne pozeráme ako keby šlo o naše dáta, ktoré sme dennodenne sťahovali z webstránky. Zaujímajú nás pri tom iba hodnoty n a m. Pre fixne zvolené N robíme 30 takýchto náhodných behov a spriemerujeme ich, aby sme dostali čo najreprezentatívnejší výsledok.Tým spôsobom vieme porovnať naše výsledky s „náhodnými“ dátami pre fixnú veľkosť množiny N, aby sme zistili, či naša funkcia nemá nejaký atypický priebeh, a teda naše predpoklady nie sú správne.

Výsledný graf je možné si pozrieť na obrázku 2. Hodnota N značí skutočnú veľkosť množiny, z ktorej náhodne vyberáme dáta, je fixné a prislúcha jednému grafu. Hodnota k je predikcia, že pri náhodných dátach, ktoré máme v danom čase k dispozícii, aká je najpravdepodobnejšia veľkosť množiny U. Je zrejmé, že k sa v čase mení aj v závislosti od konkrétnej náhodnej vzorky mení. Môžeme si všimnúť, že v priemernom prípade k pomerne rýchlo konverguje k N, čo je očakávateľné, nakoľko ak s počtom nameraných dát pôjdeme do nekonečna predpokladáme, že $k = N$.

Z grafu 2 by niekomu mohlo napadnúť, že funkcia na nami pozbieraných dátach má atypický priebeh. Vyzerá ako keby autori stránky v čase pridávali



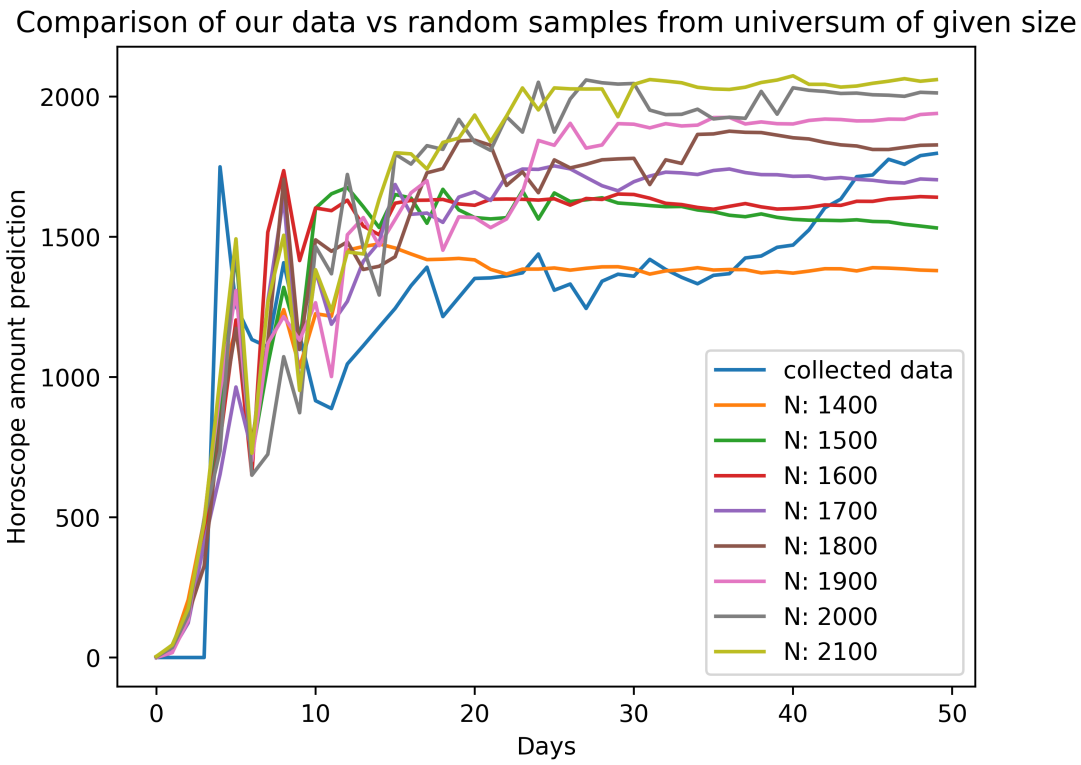
Obr. 1. Predikcia veľkosti množiny horoskopov na diva.sk v závislosti od času

nové horoskopy do ich databázy, a teda že skutočné N je rastúce. Vôbec to tak nemusí byť, nakoľko ostatné funkcie v grafe sú priemermi niekoľkých behov, teda a ide o očakávaný priebeh funkcie $f(k)$ pre rôzne konštantné N. Pokojne sa mohlo stať, že síce očakávaný priebeh funkcie vyzerá „uhladene“, ale jednotlivé behy majú pomerne veľký rozptyl.

Preto sme sa na záver nášho projektu rozhodli ešte naprogramovať vizualizáciu odchýlky funkcie určujúcej hodnotu najpravdepodobnejšieho k pre fixné $N = 1500, 1750, 2000$ v závislosti od množstva vstupných dát. Robíme to tak, že vykreslíme desať rôznych behov tejto funkcie. Na základe týchto simulácií by nám už malo byť zrejmejšie, či je priebeh našej funkcie skutočne atypický. Grafy pre konkrétne $N = 1500, 1750, 2000$ je možné si pozrieť na obrázkoch 3, 4 a 5.

4 Záver

V našom projekte sme analyzovali horoskopy zo stránky `https://diva.aktuality.sk/horoskopy/denny-horoskop/`. Všimli sme si, že identické horoskopy sa v

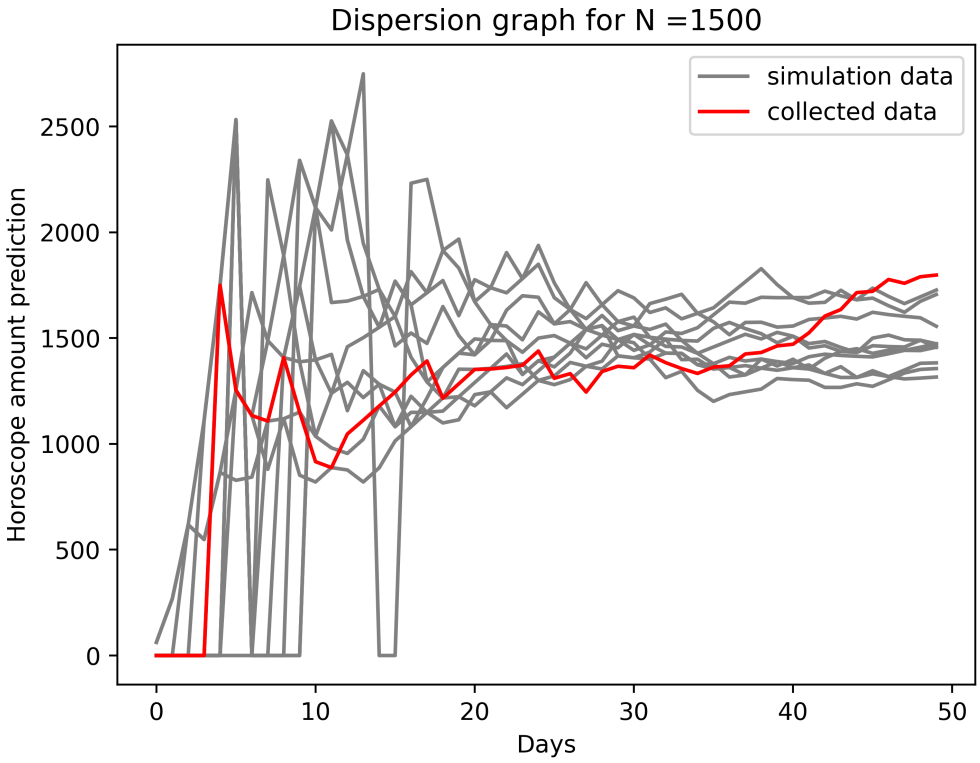


Obr. 2. Porovnanie očakávaného priebehu funkcie $f(k)$v závislosti od počtu náhodných výberov z množiny N

čase na stránke opakujú. Na základe tohto pozorovania sme sformulovali hypotézu, že autori stránky predgenerovali nejakú fixnú množinu horoskopov a následne z nej každý deň len vyberajú náhodných 12 horoskopov.

Pomocou pravdepodobnostných metód sme vypočítali najpravdepodobnejšiu veľkosť tejto množiny, ktorá je ku dnešnému dňu $k = 1798$. Súčasťou projektu je tiež porovnanie priebehu funkcie s náhodným výberom z množiny fixnej veľkosti N pomocou Monte-Carlo simulácie, a to pre rôzne N (zgenerovali sme očakávaný priebeh tejto funkcie). Následne sme pre $N = 1500, 1750, 2000$ vykreslili desať rôznych behov tejto funkcie náhodnej premennej. Cieľom bolo poukázať na možný rozptyl funkcie nakoľko funkcia na našich dátach sa správa atypicky v porovnaní s očakávaným behom.

Zrejme najférovejším záverom našich pozorovaní by bolo povedať, že na potvrdenie alebo vyvrátenie našej hypotézy nemáme ešte dostatok dát. Ak sa pozrieme na graf 4, môžeme si všimnúť, že nami zozbierané dáta zapadajú do rozptylu priebehu funkcie pre $N = 1750$, takže môžeme očakávať, že v databáze

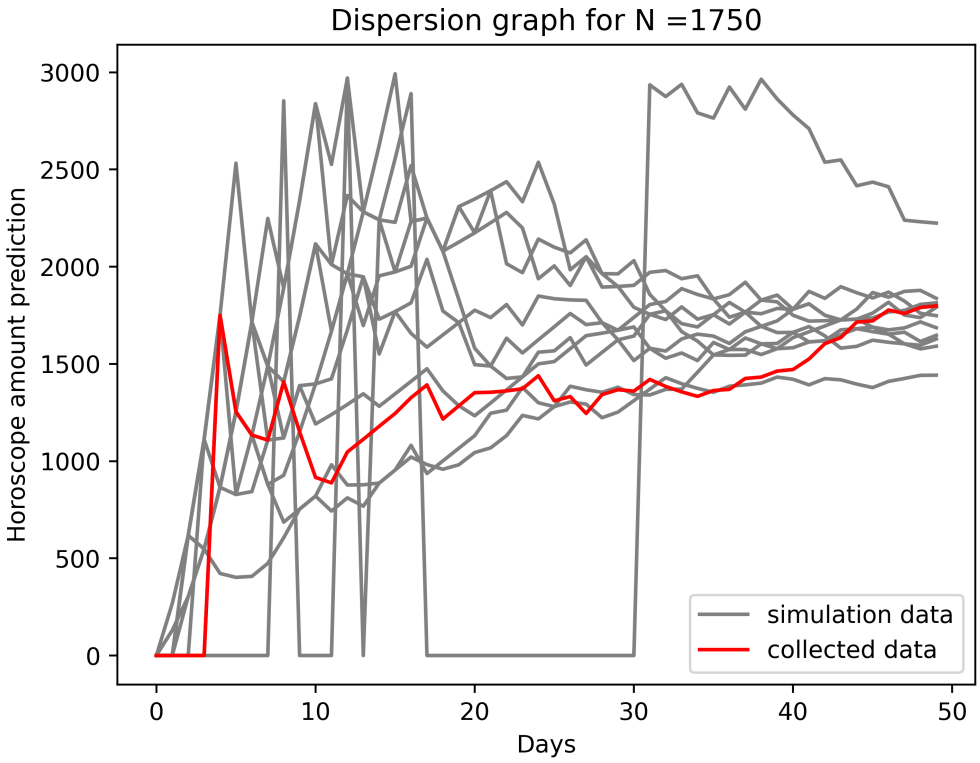


Obr. 3. Grafická vizualizácia rozptylu funkcie pre $N = 1500$

na diva.sk stránke majú okolo 1750 rôznych horoskopov (ten rozsah v ktorom sa môžeme pohybovať vie byť dosť veľký, pravdepodobne to je nejaká hodnota medzi 1500 a 2000). Súčasne ale možné vysvetlenie pre priebeh našej funkcie je aj to, že autori postupne pridávajú do svojej databázy nové horoskopy. Toto vysvetlenie sa na prvý pohľad, obzvlášť ak sa pozrieme len na graf 2 môže javiť ako najpravdepodobnejšie, avšak dovolíme si tvrdiť, že to tak nemusí byť. Pracujeme s náhodnými veličinami, a ako ukázal graf vizualizujúci odchýlku pre $N = 1750$, dobrým vysvetlením priebehu našej funkcie je aj to, že ide o bežné správanie tejto funkcie.

Osobne sa nateraz prikláňame k druhej možnosti, lebo podľa Occamovej britvy by sme sa, ak je viacero možných vysvetlení toho istého javu, mali prikloniť k tomu, ktorý poskytuje najjednoduchšie vysvetlenie. V každom prípade, na to, aby sme dospeli k presnejším záverom by bolo najsprávnejšie zbierať dáta ešte niekoľko mesiacov.

Poznamenajme na záver, že z nami vyzbieraných dát môžeme s istotou tvrdiť, že horoskopy na stránke `https://diva.aktuality.sk/horoskopy/denny-horoskop/`



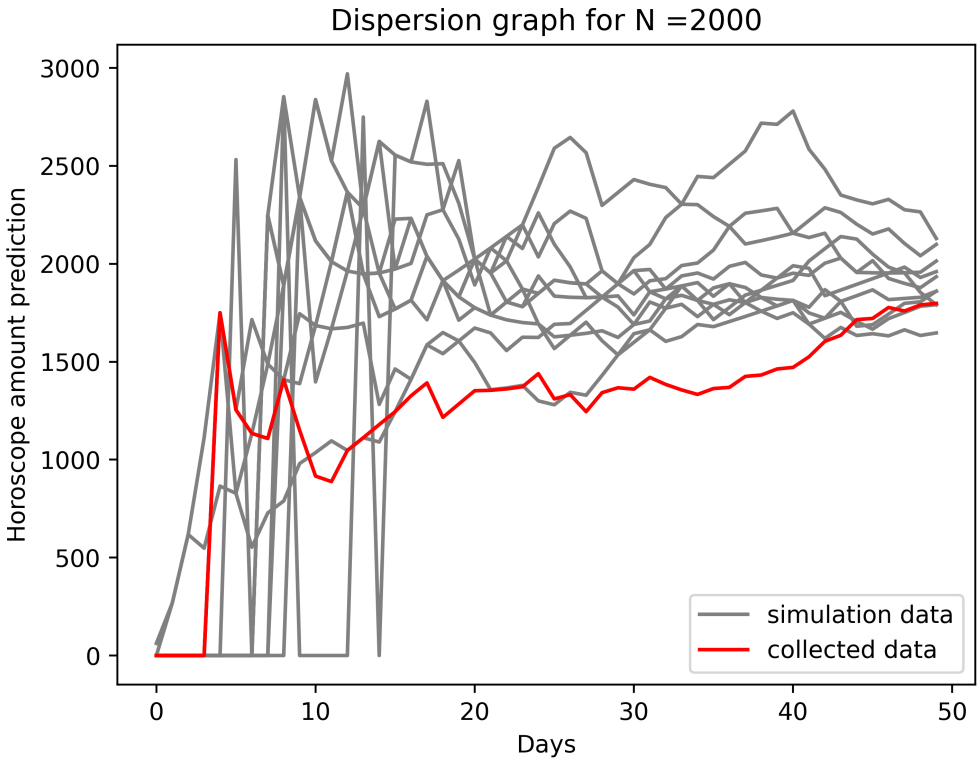
Obr. 4. Grafická vizualizácia rozptylu funkcie pre $N = 1750$

sa celkom pravidelne opakujú, čo výrazne podkopáva dôveryhodnosť, ale aj profesionalitu tejto stránky. Pre ilustratívnosť, jeden horoskop sa počas dvoch mesiacov objavil na stránke dokopy až štyrikrát.

5 Použité nástroje a diskusia

V poslednej kapitole zhrnieme nami použité metódy získavania a spracovania dát a ich odôvodnenie. Hlavnými časťami našej práce sú dva programy *crawler.py* a *main.py*. Úlohou crawlera je každý deň ráno o siedmej (stará sa o to príkaz *crontab*) prísť na stránku `https://diva.aktuality.sk/horoskopy/denny-horoskop/` a stiahnuť denné horoskopy pre jednotlivé znamenia do lokálnej sqlite3 databázy.

Jeden riadok v tejto databáze obsahuje nasledovné položky: *id záznamu*, *text horoskopu*, *znamenie horoskopu* a *deň, kedy bol horoskop stiahnutý*, vo formáte YYYY.MM.DD, pre ľahké triedenie dátumov.



Obr. 5. Grafická vizualizácia rozptylu funkcie pre $N = 2000$

Okrem databázy sme tiež pre istotu záznamy sťahovali aj do textového súbora ako zálohu, nakoľko nové dáta by už nebolo možné získať a ohrozilo by to priebeh nášho projektu, keby sa náhodou niečo s databázou stalo.

Program *crawler.py* sme na začiatku projektu oddebuggovali a manuálne sme pre pár dní overili, že záznamy zapísané v databáze sú rozumné. Následne už nerobíme automatizované kontroly dát. Rozhodli sme sa kvalite dát dôverovať, nakoľko si záznamy v databáze vytvárame sami, pričom kvalita samotného textu horoskopu, ktorý sťahujeme z webstránky nás až tak nezaujíma, zaujíma nás hlavne to, či sa medzi sebou zhodujú.

Program *main.py* má za úlohu stiahnuť potrebné dáta z databázy, spraviť popísanú Monte-Carlo simuláciu, vypočítať priebeh funkcií a na záver pomocou Pyplotu vykresliť priebeh funkcií do obrázkov, ktoré môžeme vidieť v našom reporte. Ako sme už spomínali hľadanie optimálneho k pre maximálnu hodnotu funkcie $\binom{k}{m}/k^n$ má za úlohu program v jazyku c++, ktorý main.py používa ako externú knižnicu.

Všetky programy sú spustiteľné bez akéhokoľvek ďalšieho používateľského vstupu.

Celkovo považujem nami zvolené nástroje za vhodne zvolené a dobré. Počas projektu sme sa v jednom momente rozhodli preprogramovať funkciu na výpočet $\binom{k}{m}/k^n$ do jazyka c++, čo vnímam ako pozitívne rozhodnutie, nakoľko to signifikantne urýchlilo naše výpočty. V súčasnosti mi nenapadá, čo by som na projekte robil inak, skôr by som chcel oceniť vyučujúcich tohto predmetu, že voľbou jednotlivých tém mi rozšírili v tom, aké nástroje sa dajú používať. V podstate len pomocou vecí, ktoré sme preberali na prednáškach som vedel bez väčších problémov naprogramovať všetky časti môjho projektu .