

Velká geodata, lokalizace mobilních telefonů a zkušenosti ze světa

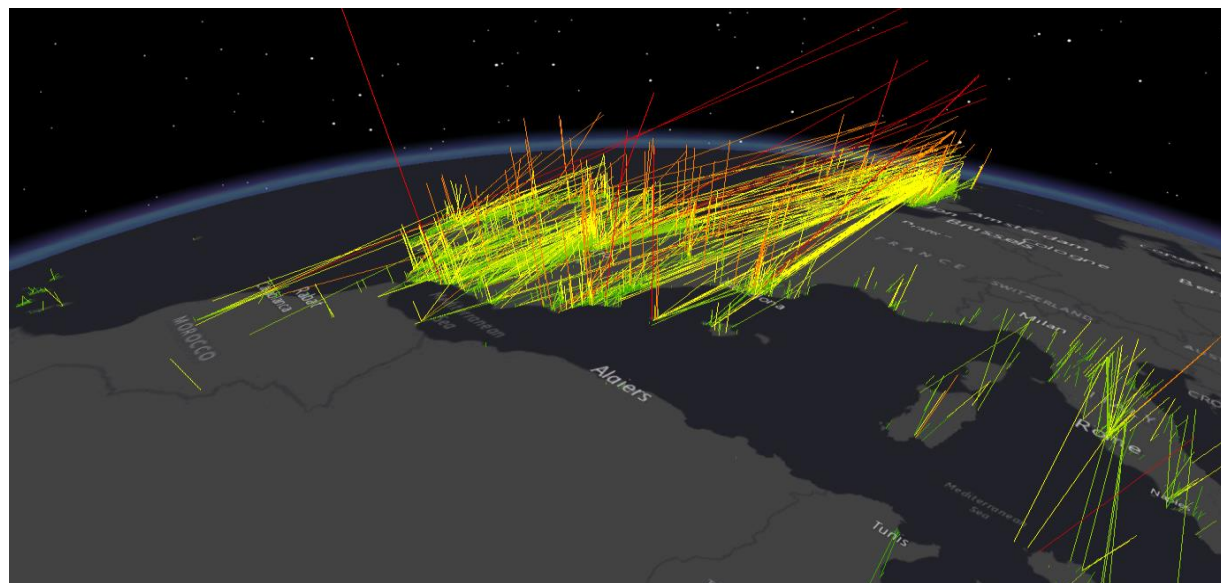
Jiří Horák

VŠB-TU Ostrava, Institut geoinformatiky

- Hlavní rysy Big data:
 - Volume - velký objem, který lze jen obtížně zpracovat tradičními způsoby a který vyžaduje nové formy ukládání, správy a zpracování dat,
 - Velocity - rychlost sběru a přenosu dat, ale také rychlost potřeby jejich zpracování. Hledají se nové algoritmy efektivní analýzy proudů dat.
 - Variety - heterogenita dat, která je spojena s využíváním široké škály zdrojů dat a jejich rostoucí nekonzistencí
 - Value – hodnota, která se vztahuje k procesu analýzy cenných informací z velkých dat. Vznikají nové oblasti použití cenných informací.
 - Veracity – věrohodnost, velký význam přesnosti informací a vypovídací hodnotě. Popis kvality dat se udržuje správou metadat. Řeší se zde i otázky soukromí a právní aspekty.
- Big geodata přidávají:
 - Visualization – nové postupy využívající vizualizaci pro nalezení vzorů (resp. odlehlé hodnoty, shluky), formulaci nových hypotéz a postupů zpracování.
 - Visibility – snadné sdílení dat pomocí cloudových technologií.

- Masivní měření – lidar, družicová data, senzorové sítě apod.
- Masivní objemy dat ze sociálních sítí – Twitter
- Individuální monitoring – pohyb, aktivity (human mobility pattern) atd. – sledování skupiny (flotily) lidí, aut apod.
- Wifi konektivita
- Často součást „smart city“ konceptu
- Data z lokalizace mobilních telefonů, resp. SIM karet

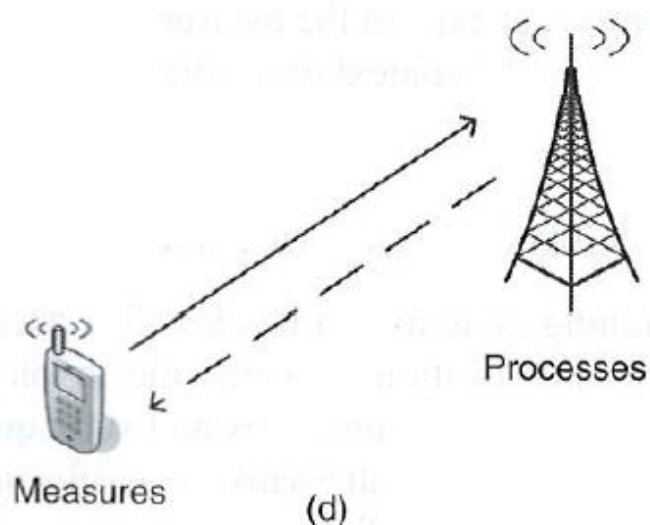
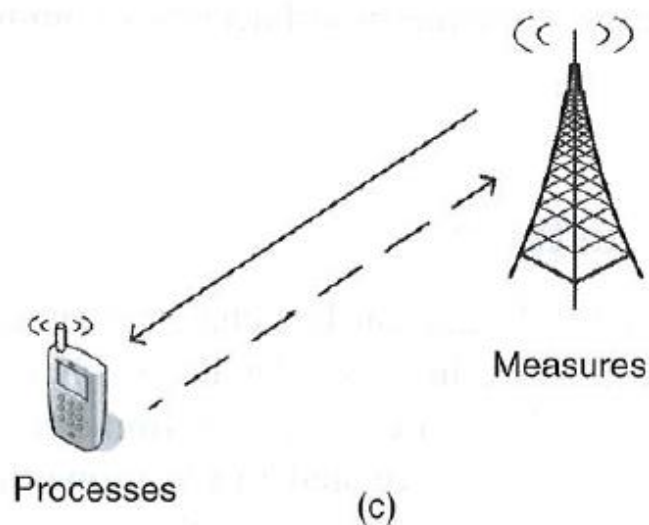
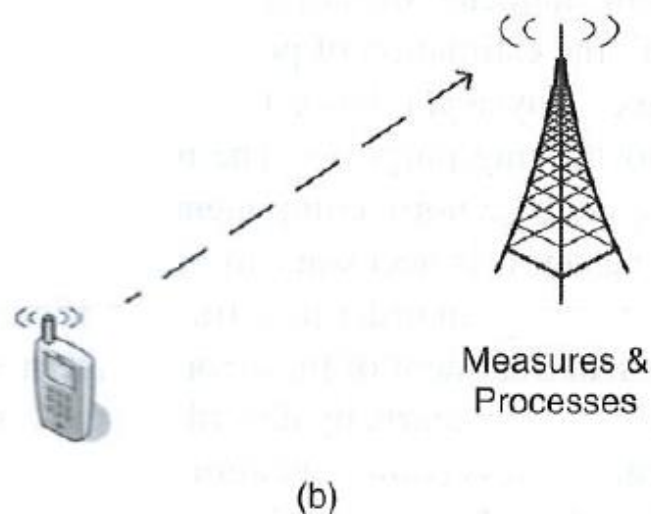
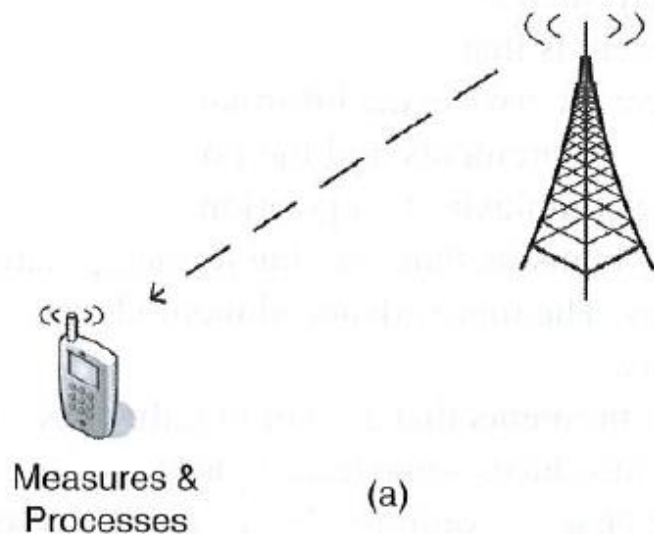
David Kocich –
časoprostorové
pohyby uživatelů
Twitter



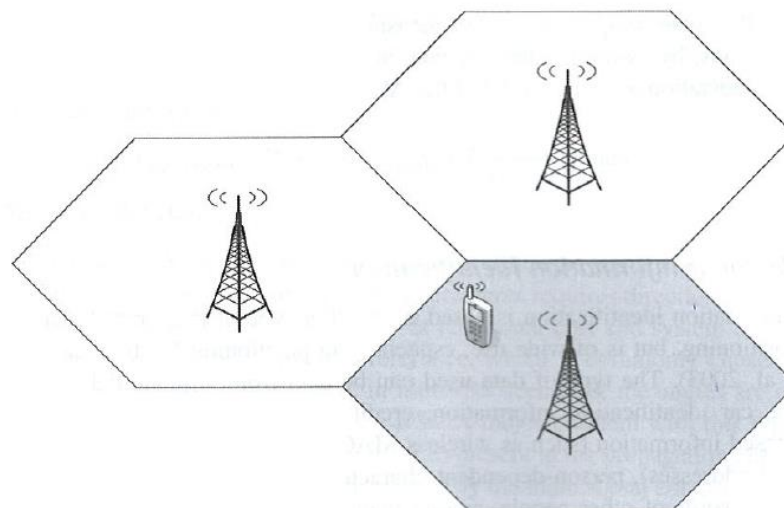
- data mining (dolování informací z obrovského množství dat)
- nástroje business intelligence (reporting pro manažerské rozhodování)
- učící se stroje
- učící se systémy vyhodnocující data v reálném čase
- neuronové sítě, zejména DNN
- Např. pro klasifikaci objektů v rychle jedoucím autě/UAV
- vizualizace analytických dat – viz prezentace Igor Ivan

- Možnosti mobilní lokalizace ve venkovním prostředí
- Základní přístupy:
 - mobile-assisted GPS-free positioning (aktivní)
 - network-based GPS-free positioning (pasivní)
- typy měření – Cell ID, síla signálu, úhel
- lokalizační techniky – cell ID (s průměrem), triangulace a alterace, hyperbolická lokalizace, protínání vpřed z úhlů, databázová korelační technika (fingerprinting), hybridní metody
- problémy s non-line-of-sight NLOS, vícecestné slábnutí, zastínění
- Více prezentace čtvrtek večer „Rychlý internet jako budoucnost lokalizace mobilních zařízení“

- 4 základní topologie – self-positioning, b - vzdálená lokalizace, c - nepřímá self-positioning, d - nepřímá vzdálená lokalizace



- Každá stanice je identifikována kódem země (MCC Mobile country code, kódem sítě (MNC Mobile network code), kódem oblasti (LAC local area code) a ID buňky (cell ID) – ne vždy jsou ale všechny kódy k dispozici, někdy je mobilní síť rozdělena na sekce (např. 3)
- mobil se připojí k jedné BTS, identifikované svým ID
- Základní způsob lokalizace - poloha mobilu je nahrazena polohou identifikované BTS



- Geometrie buněk kolem BTS se nebere v úvahu (často jsou navíc antény BTS na okraji sektorů buněk)
- Vylepšení – na základě tvaru buňky umístění mobilu do virtuálního těžiště, rozdělení buněk do segmentů podle dalších měření a překrývání BTS

- soubory CDR (call detail records) („zbytková data“)
- záznamy, které síťová infrastruktura (platební logger) generuje pro každou aktivitu uživatele mobilu – iniciace hovoru, posílání SMS, přístup na Internet.
- CDR obsahuje ID uživatele, ID provádějící BTS, datum a čas.
- anonymizované informace od telefonních operátorů, informující o počtu aktivních SIM karet v daném území
- Agregace v čase - telekomunikační aktivita v erlanzích (E)
- Agregace v území - otázka vhodné územní jednotky

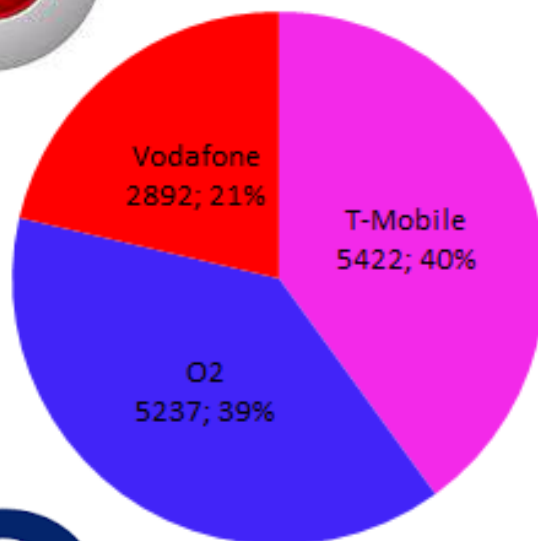
- Ve srovnání s jinými zdroji poskytujícími podobný výsledek (záznamy GPS, WiFi):
 - nízká cena 😊
 - připravená infrastruktura
 - real-time data
 - studie různě velkých území (i celý stát)
 - lze zmapovat rozsáhlé jevy, i když se těžko vysvětlují
 - krátký čas první lokalizace (ve srovnání s GNSS)
- Problémy individuálních záznamů – často to bývá specifická skupina dobrovolníků, typicky studenti v univerzitním kampusu
- Problémy GPS monitoringu lidí – fragmentace cest (ztráta signálu, pohyb v budovách). Např. v Jahromi (2016) více než polovina trajektorií trvala méně než hodinu.

- Problém stability sítě v čase - ne všechny BTS jsou stále aktivní, zapojení nových.
- Systematická odchylka - hustší síť je ve více urbanizovaných místech, významnějších místech nebo bohatších místech s vyšší penetrací mobilů
- Zahrnutí využití území (dasymetrie). Oblasti bez pokrytí signálem vykazují nulovou mobilitu či výskyt osob.
- 1 hovor může být předán jiné BTS v rámci vyrovnávání provozu (cell balancing, kvůli kapacitě sítě a síle signálu), aniž by se uživatel pohnul. To se stává v oblastech s vysokou hustotou mobilní sítě. Občas chronické přepínání mezi BTS.
- Neautorizovaný přístup k mobilnímu telefonu a komunikace ve stejný čas jako uživatel, roboti
- Nejsou údaje od všech operátorů – může vést k sociální systematické odchylce, kdy místní populace používá jiné zastoupení mobilů a zejména jejich operátorů než je očekávaný průměr (model)

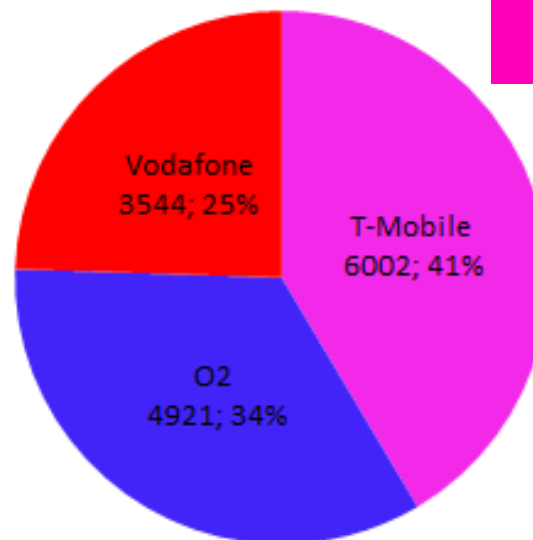


Počty SIM karet v tisících:

2008



2016

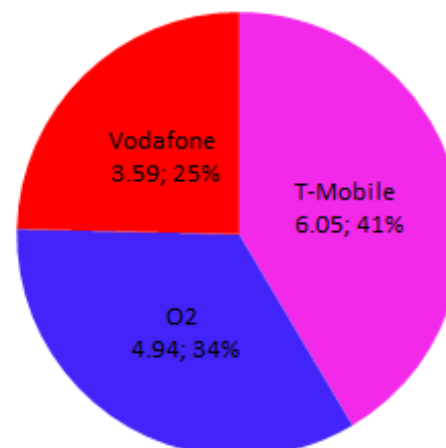


T-Mobile

T-Mobile
O2
Vodafone

O₂

2016

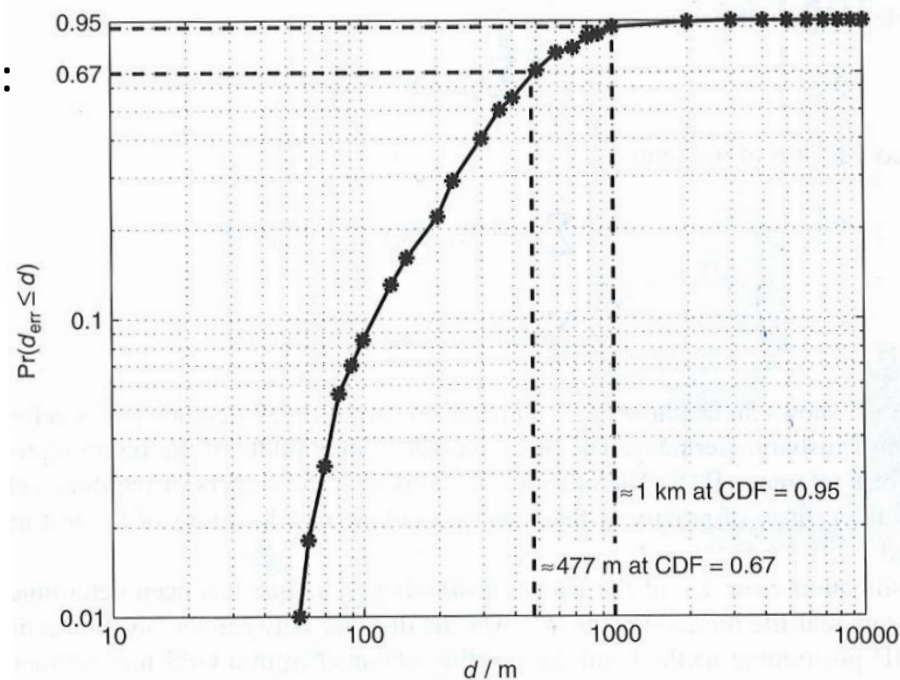


Počty zákazníků v milionech:

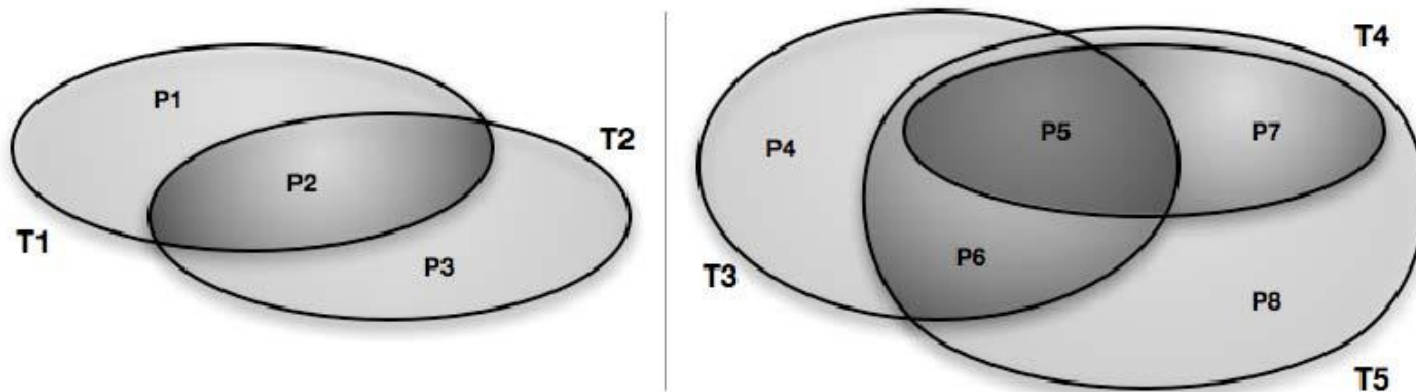
- Standardní určování hranic sektorů je podle Thiessenových polygonů pro BTS a přidělení nejbližší BTS.
- omezené prostorové rozlišení definované velikostí plochy BTS. Relativní přesnost 15-20% poloměru buňky. Obecně to bývá od několika stovek m² do několika km².
- Jahromi (2016): Milán 2012-3: průměrná servisní plocha BTS 3 km² a jen 30% BTS mělo méně než 1 km².

ohodnocení polohových chyb v Duisburgu:

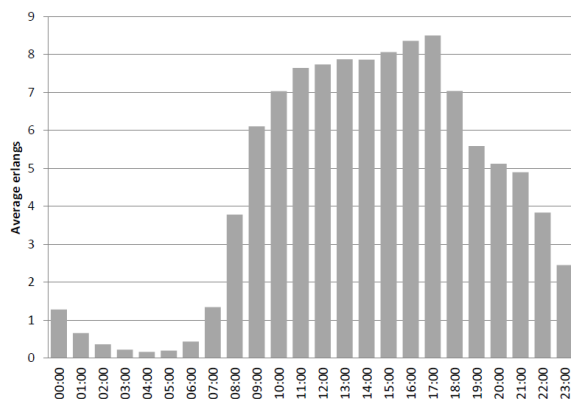
- 50% lokalizace s chybou do 356 m
- 67 % lokalizace s chybou do 447 m
- 95% lokalizace s chybou do 1000 m



- Problém příliš husté sítě a překrývání Girardin et al. (2009)
- Vypočítali radiovou mapu segmentů, generovaných překrytím sektorů jednotlivých BTS.
- Okumura-Hata model - hranice sektoru se odvozují z polohy, výšky, azimutu, Effective Isotropic Radiated Power (EIRP), typ antény a frekvence vysílání + analýza fyzického prostředí okolo (např. přítomnost řeky). Z překryvů sektorů nakonec vytvořili 9900 nepřekrývajících se segmentů.
- aktivita v každé oblasti zájmu odhadnuta pomocí relativní váhy příspěvku každého sektoru do statistiky telefonátů v každém segmentu a jeho překryvu s plochou zájmu . Konkrétně – příspěvek každé BTS do segmentu je nepřímě úměrný počtu vysílačů pokrývajících daný segment. Pokud se tedy překrývají 2 BTS, hustota aktivity v části P1 od stanice T1 bude 2x větší než hustota aktivity pro segment P2 (kde se překrývá s T2). Tj. $d(P1)=2*d(P2)$. P5 bude mít třetinovou hustotu aktivity od T3 proti P4. Tedy $d(P4)=2*d(P6)= 3*d(P5)$. Ve stejném poměru se tedy má rozdělit počet na BTS mezi jednotlivé segmenty.



- Omezené časové rozlišení – typicky CDR obsahuje aktivity uživatele, tj. musí být aktivní na mobilu.
- Ne vždy je k dispozici i záznam internetových aktivit (což je ale potřebné)
- registrují se aktivity na mobilu, tj. více se budou pohybovat ty osoby, které více používají mobil (i když mají ve skutečnosti stejné cesty). Bohatší lidé (více používající mobil) budou falešně vykazovat větší mobilitu.
- Potřebný periodický update – co 30 min zavolá síť váš mobil, takže nejen aktivní SIM karty. Pak jsou cesty delší než 30 min dobře sledovatelné.
- sledování časových řezů a ne kontinuální sledování chování
- Jahromi (2016): pro sledování doby pobytu v každé buňce používají předpoklad, že u SMS a internetových aktivit je před každým a po každém záznamu 30 minut na tomto místě, v případě hovoru se k tomu ještě připočte délka hovoru



Střed Amsterdamu, 2010, erlangy



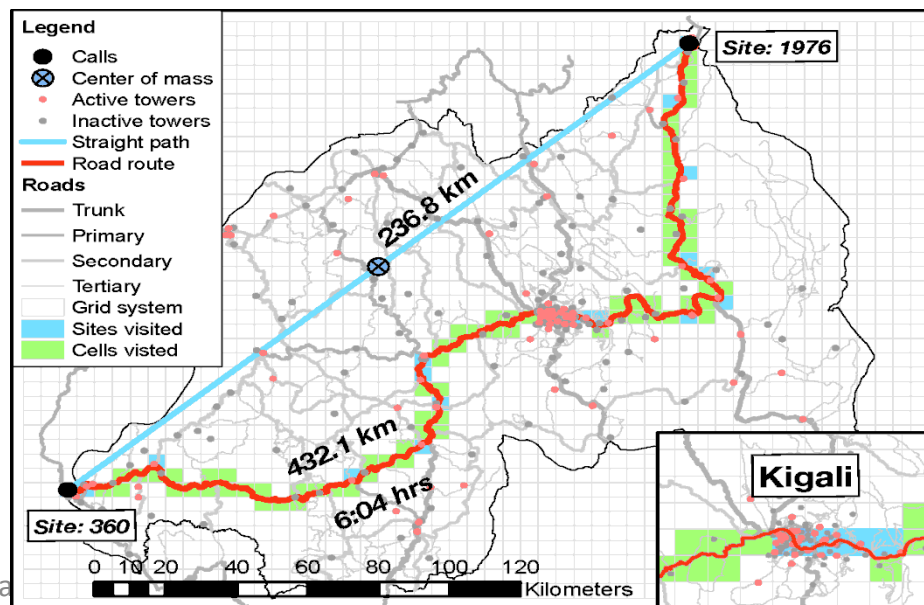
Město Shenzhen, 2012: aktivní zjišťování SIM

- **Lokální intenzita:**
- Prostý počet
- Comparative Relative Strength (CRS) – aktivita v daném místě vůči celkové aktivitě ve městě, Z-skóre
- Poměr hovorů lokálních/návštěvníků
- RR Relevance Ratio pro POI P a uživatele u:

$$RR(P, u) = d_{\text{visit}}(P, u) / d_{\text{total}}(u)$$

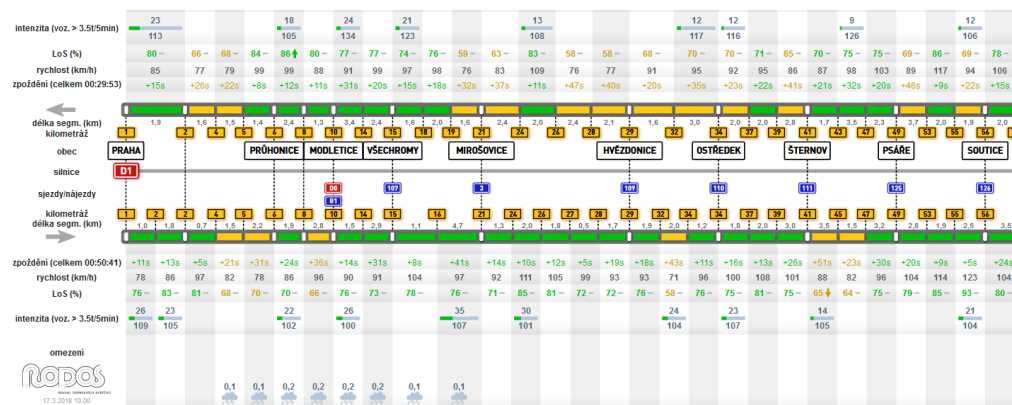
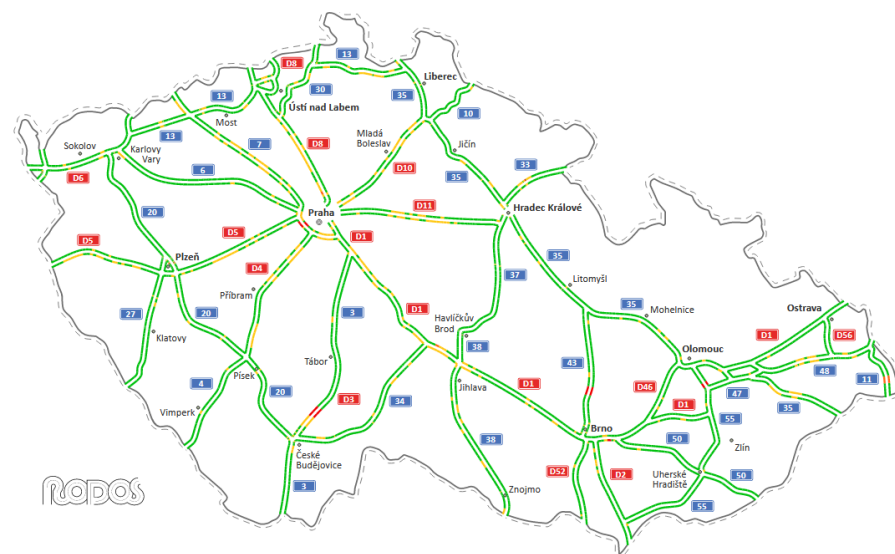
kde $d_{\text{visit}}(P, u)$ je počet dnů, kdy bylo POI P navštíveno uživatelem u
 $d_{\text{total}}(u)$ počet dnů zaznamenaných pro u

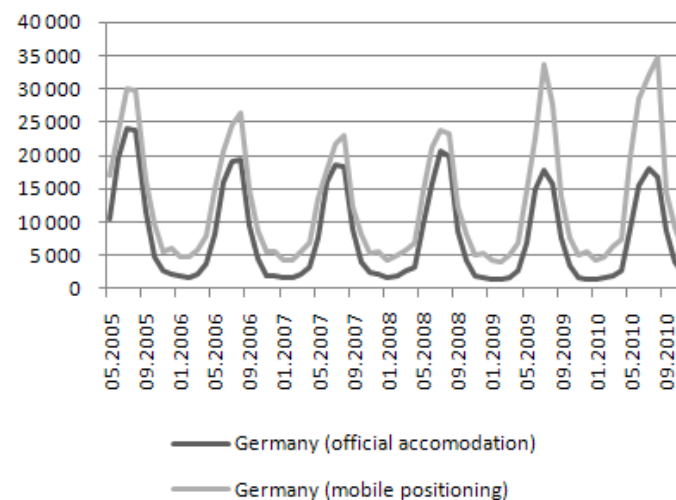
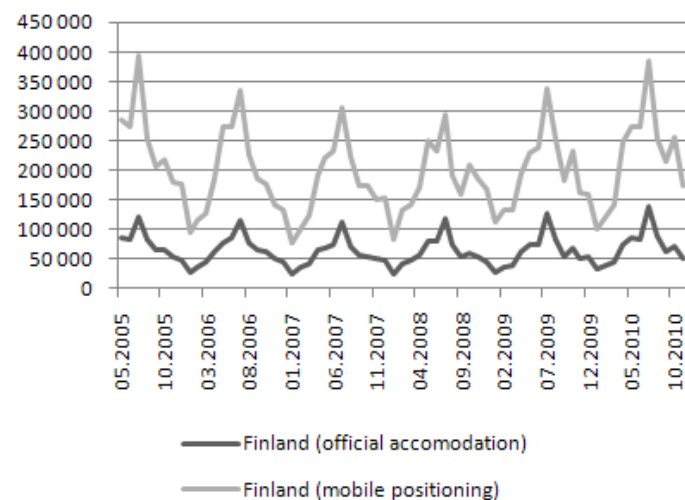
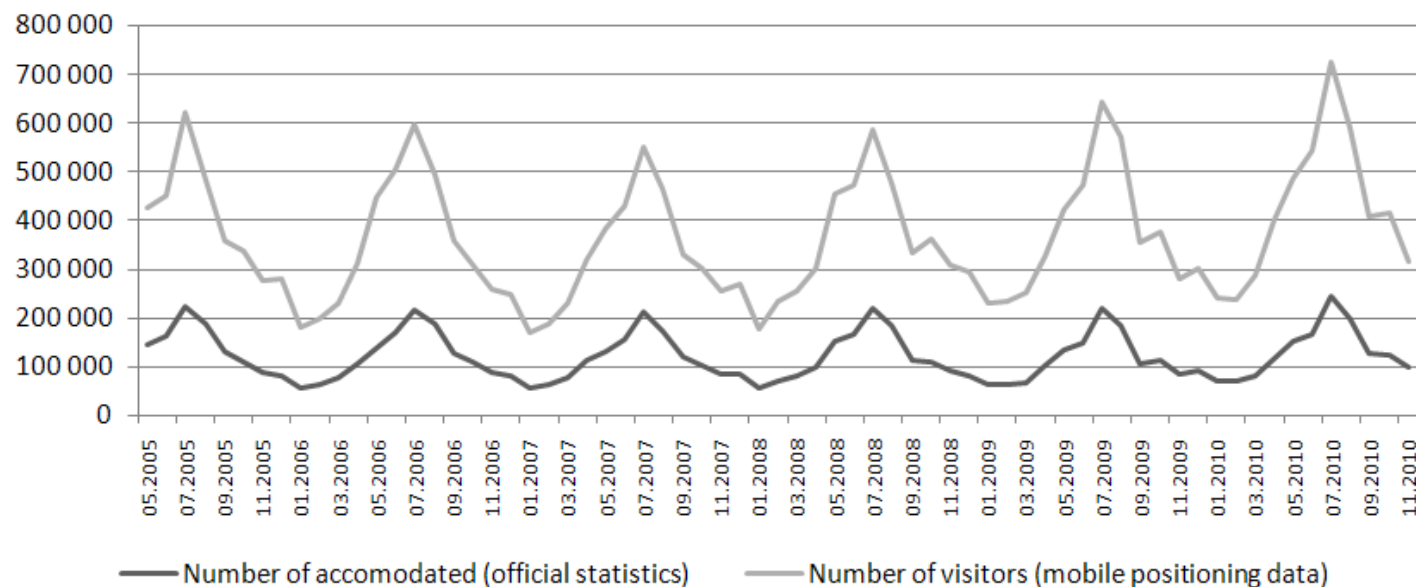
- **Sledování mobility osob** (v daném časovém intervalu – hodiny, dny, měsíce):
 - Počet použitých BTS (number of tower used, NTU) – počet stanic, ze kterých uživatel použil mobil
 - Přímková vzdálenost (distance traveled-straight line, DT-SL, resp. průměrná cestovní vzdálenost) – součet přímkových spojnic mezi BTS v posloupnosti, odpovídající použití mobilu
 - Maximální vzdálenost (maximum distance traveled, MDT) – maximální přímková vzdálenost mezi 2 použitými BTS
 - Poloměr otáčení (radius of gyration, RoG) – nejvíce používaná míra – odpovídá střední polohové chybě kolem těžiště volání. Nejdříve se určí těžiště ze všech použitých BTS a následně se spočítá odmocnina ze sumy čtverců přímkových vzdáleností mezi jednotlivými BTS a těžištěm



- **Sledování mobility osob** (v daném časovém intervalu) – **doporučený postup**:
- Použití pravidelné čtvercové sítě se stejnou velikostí buňky v celém území (přiměřená k velikosti BTS), sledují se jen buňky s nejméně 1 aktivní BTS.
- Pohyb se registruje jen v případě přechodu z 1 buňky do druhé.
- výhoda ve standardizaci územní jednotky s jinými zdroji dat
- Počet aktivních buněk se sice s časem mění, ale ne tak výrazně jako počet BTS
- Použití informace o cestě mezi buňkami, kudy se musel uživatel pohybovat.
- Metriky – sledování frekvence mobility a rozsahu mobility:
 1. **Počet cest** – kolikrát uživatel měl hovor z následující buňky.
 2. **Počet navštívených buněk** - započítávají se všechny buňky podél trajektorie mezi 2 místy s hovorem.
 3. **Poměr navštívených míst** – poměr mezi počtem navštívených míst a celkovým počtem míst na dané trajektorii ve sledovaném časovém intervalu.
 4. **Procestovaná vzdálenost** – délka nejkratší cesty po silnici mezi 2 místy.
 5. **Procestovaný čas** – odhad času cesty po nejkratší trase mezi 2 místy.
 6. **Procestované buňky** – počet buněk gridu, protnutých nejkratší trasou mezi 2 místy.
- Williams et al. (2015)

- Sledování distribuce a struktury populace v území podle času (dynamická populace), sledování časových cyklů (den, týden apod.)
- Krizový management, např. optimalizace nasazení SaP podle rozložení popul.
- Dopravní analýzy, sledování dopravní proudů a kongescí v reálném čase
- RODOS – kombinace dat z flotily desítek tisíc „plovoucích vozidel“ a CDR dat
- Dynamika turistů a jejich statistika - počet a časová struktura návštěvnosti hlavních turistických cílů v Praze pro Czechtourism, projekty IPR, Estonsko, Manhattan atd.
- Charakteristiky chování zákazníků (CE traffic)
- Analýza lokality (CE traffic)

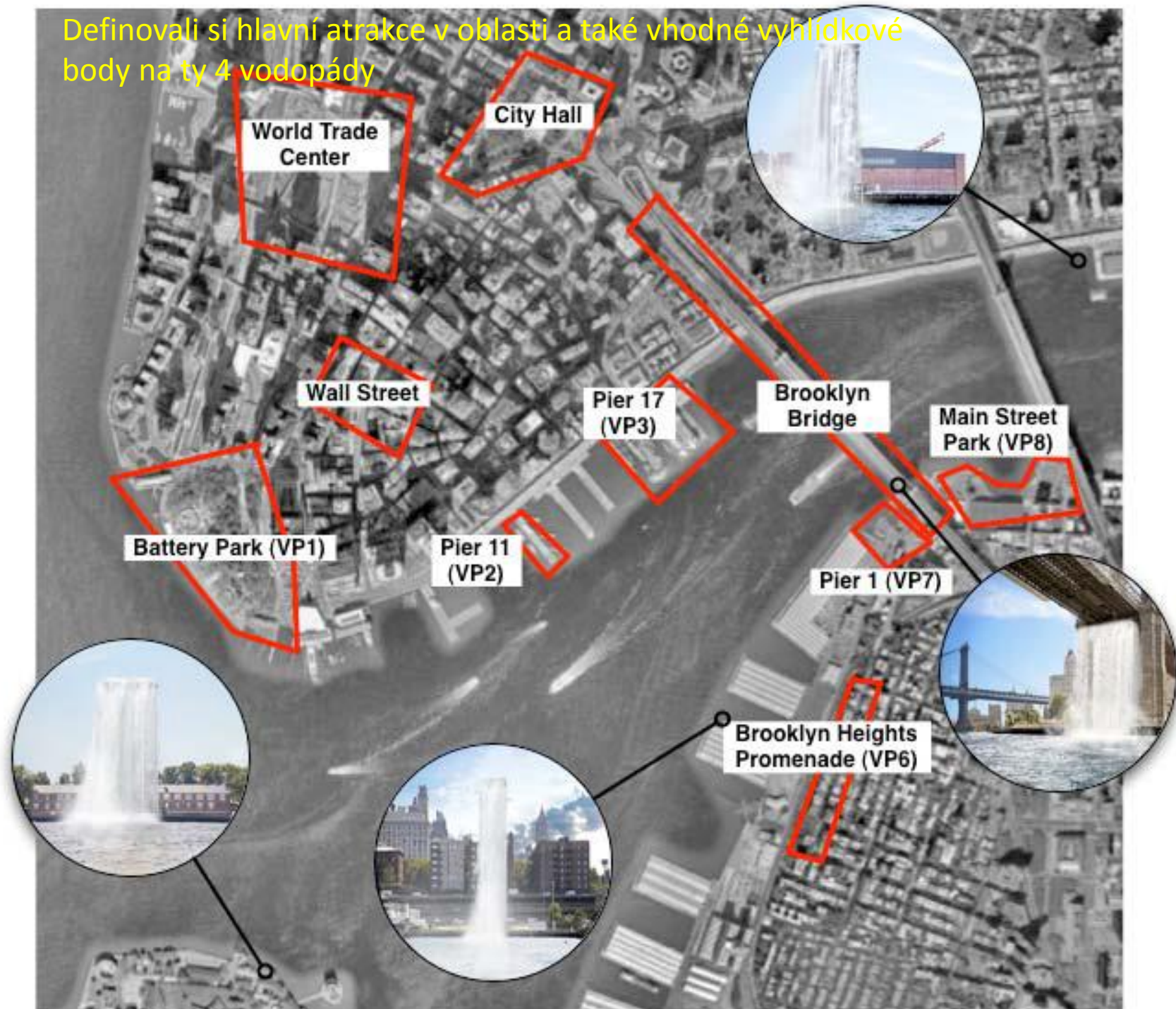




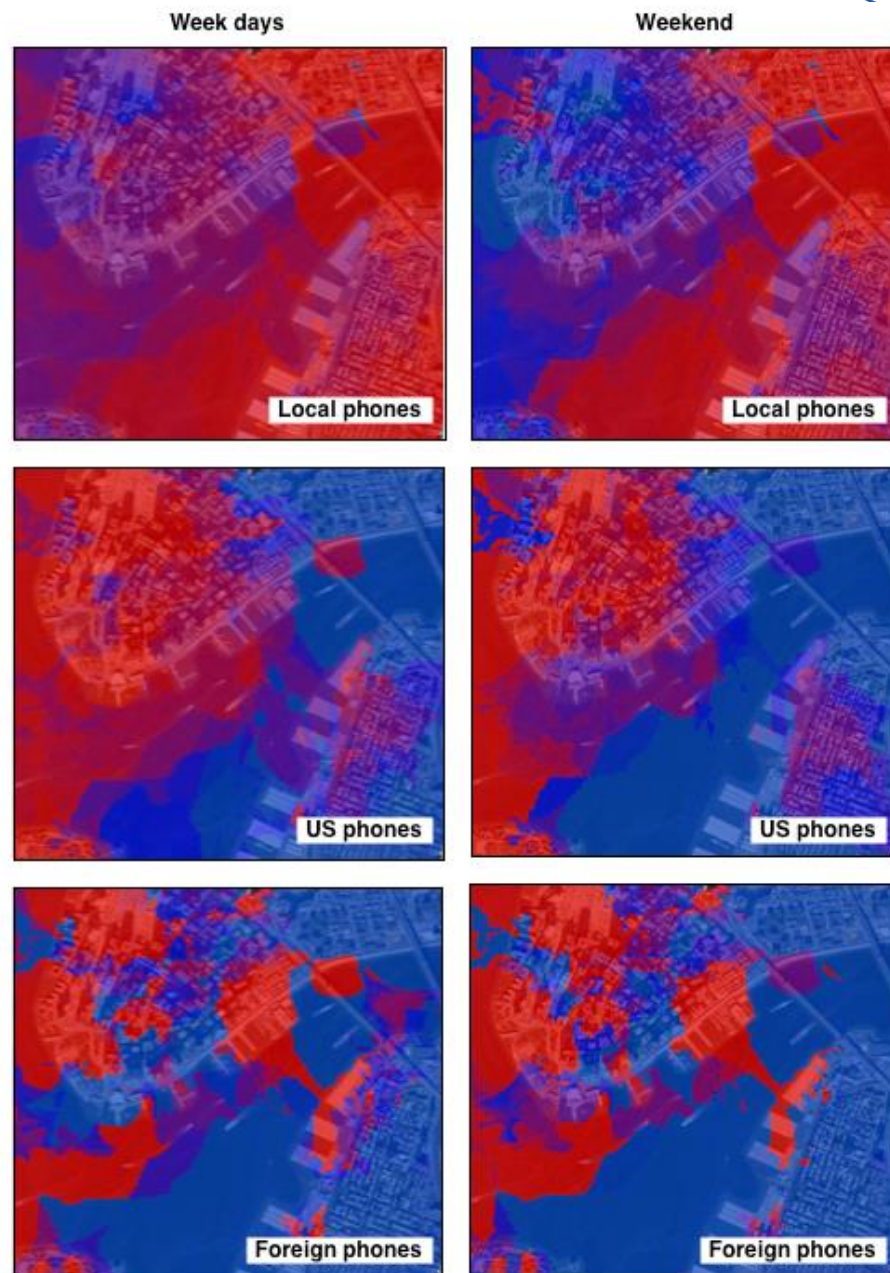
Počet ubytovaných x počet návštěvníků z CDR (Estonsko, Finsko, Německo)

- Girardin (2009)
- Data z lokalizace mobilních telefonů (AT&T síť) a Flickr
- sledování uměleckého projektu New York City Waterfalls (26.6.-13.10.2008 , 79200 nových návštěvníků přibylo, 69 mil. USD ekonomický dopad)
- Agregované údaje o počtu volání, SMS a souhrnné aktivitě síťového provozu na anténách sítě AT&T za každou hodinu, rozsah 8/2007 až 8/2008, oblast dolní Manhattan a West Brooklyn.
- Obsahují i kód oblasti, kde je domov registrovaného vlastníka mobilu (cizinci pak jen stát).
- Kvůli intenzitě překrývání sítě použili rozdělení do nepřekrývajících segmentů s relativními podíly vah.

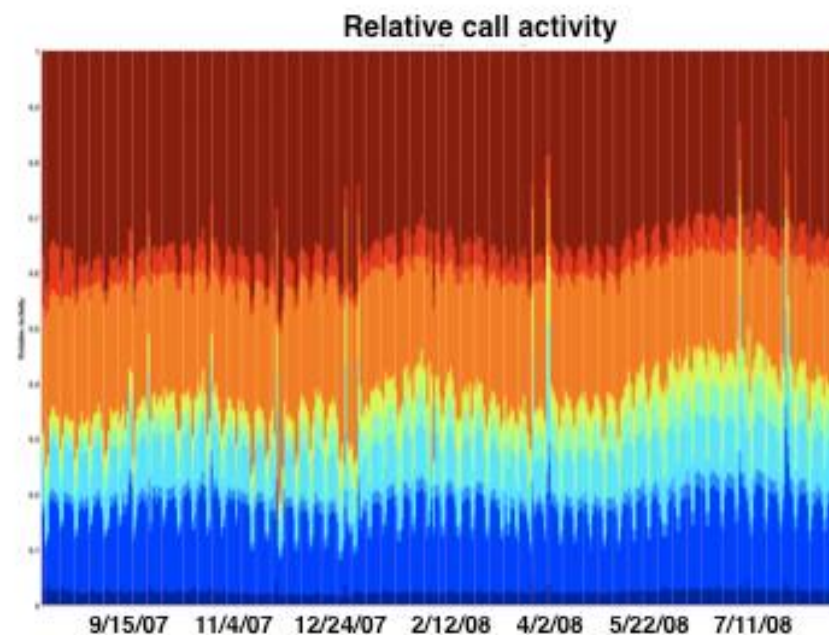
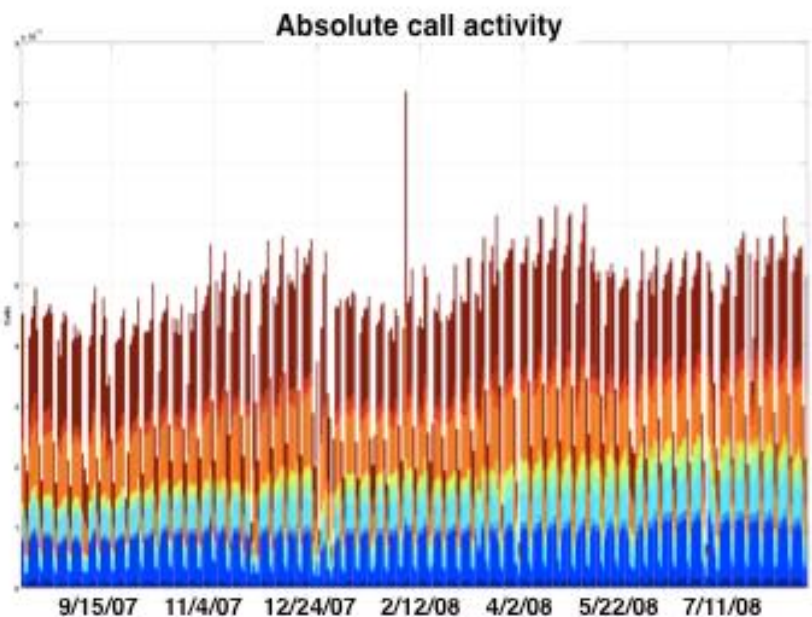
Definovali si hlavní atrakce v oblasti a také vhodné vyhlídkové body na ty 4 vodopády



- Nejdříve data jen za týden (8/2008)
- analýza rozdělila uživatele do 3 skupin – domácí z NY, ostatní z USA a cizinci.
- Výsledky – rozdíly v četnostech (je vidět, že místní opouštějí o víkendu dolní Manhattan):



vývoj aktivity během roku pomocí 2 grafů
(absolutní počty a relativní podíl ze všech sledovaných míst)



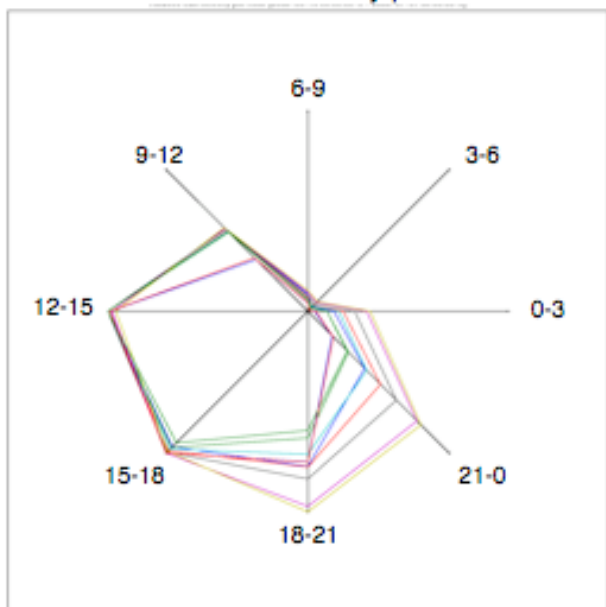
VP1 - Battery Park
VP2 - Pier 11
VP3 - Pier 17

VP6 - Brooklyn Heights
VP7 - Pier 1
VP8 - Main Street Park

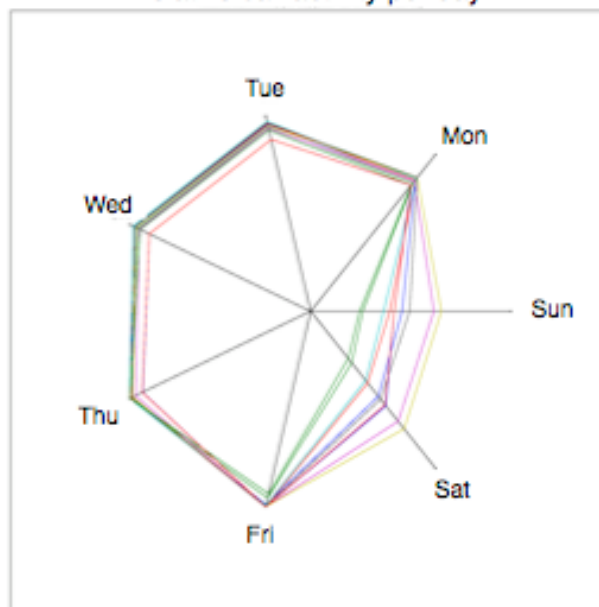
Brooklyn Bridge
City Hall
Wall Street
World Trade Center

časové vzory aktivity během dne, týdne, roku:

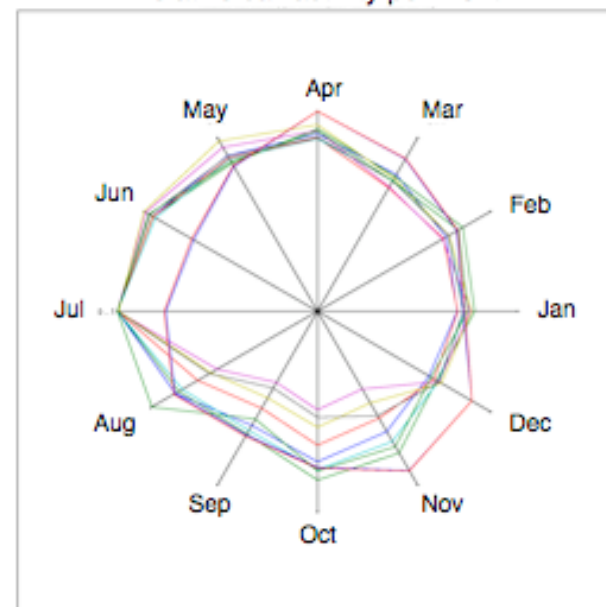
Relative call activity per hour



Relative call activity per day



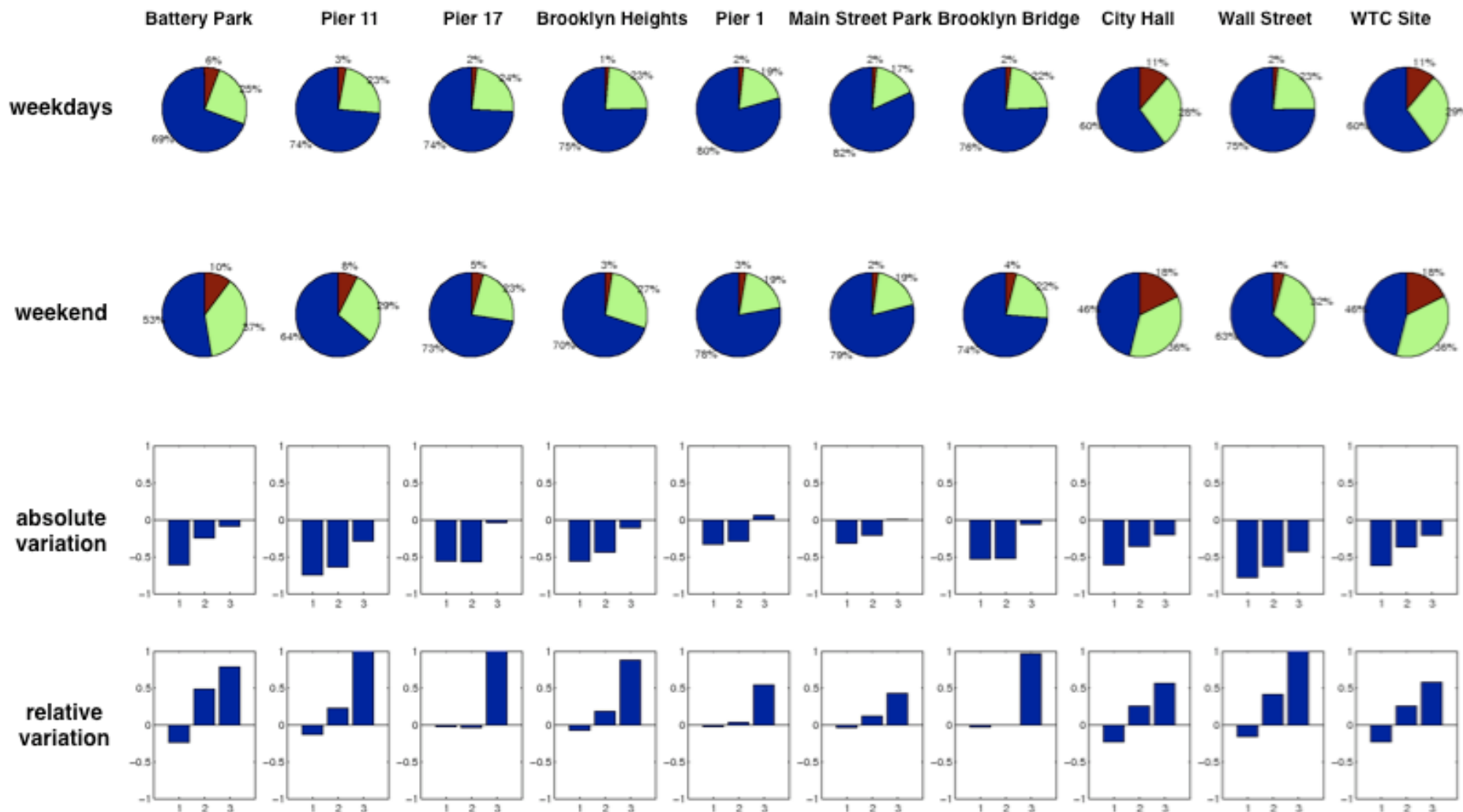
Relative call activity per month



- některá místa velký pokles (30%) během víkendu

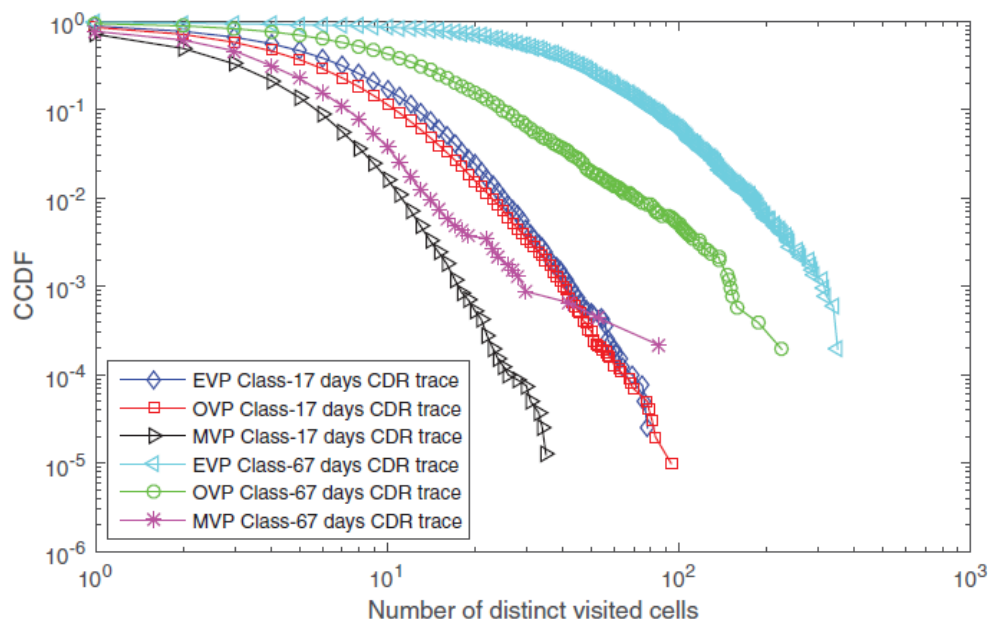
- ukazatele atraktivnosti:
- Comparative Relative Strength (CRS) – aktivita v daném místě vůči celkové aktivitě ve městě – 3 formy:
 - Přítomnost fotografa (Flickr)
 - Počet hovorů (mobily)
 - Srovnání hovorů lokálních uživatelů vůči návštěvníkům
- Absolutní počty hovorů v jednotlivých místech byly z důvodu ochrany osob upraveny (rozmazány) náhodným faktorem
- Počet hovorů – potvrdil se nárůst atraktivnosti míst, odkud bylo vidět na vodopády (o 39%, 2x apod.), ale hlavně pro cizince (návštěvníci z USA jeví malý zájem)

Poměr hovorů lokálních/návštěvníků – srovnání počtu hovorů prac.den x víkend pro jednotlivá místa absolutně a relativně (CRS) (modře místní, zeleně USA, červeně cizinci)

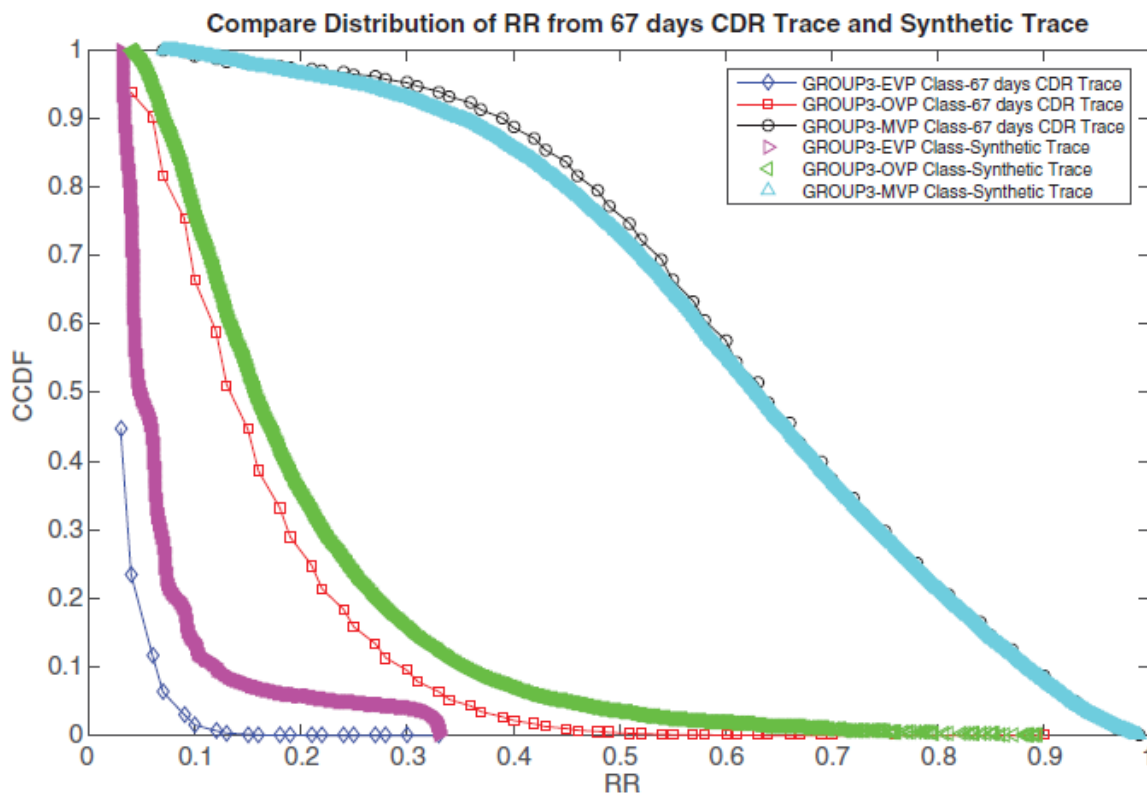


- Jahromi et al., 2016: sledování pohybu lidí - jejich rozdělení do skupin podle zastoupení různě navštěvovaných míst a hodnocení zastoupení cest mezi různě navštěvovanými místy
- 2 datové soubory CDR se záznamy všech aktivit mobilních telefonů (hovory, SMS, internetové aktivity) z evidence mobilního operátora v Milánu. CDR-17 pro 17 dní 1.-17.5.2013 pro Milán a okolí (543 tisíc uživatelů), CDR-67 pro období 26.3.-31.5.2012 pro město Milán (17 tisíc)
- Porovnání s GPS loggery (Peking) a WiFi trasováním (Dartmouth univerzita)
- Použili RR of POI
- 3 skupiny POI podle návštěvnosti – MVP (Mostly), OVP (Occasion.), EVP (Exception.)
- seskupují POI pro každého uživatele zvlášť, klasifikace uživatelů do 3 skupin, sledují počet typů POI dle návštěvnosti ve 3.skupině (nejméně usedlá)

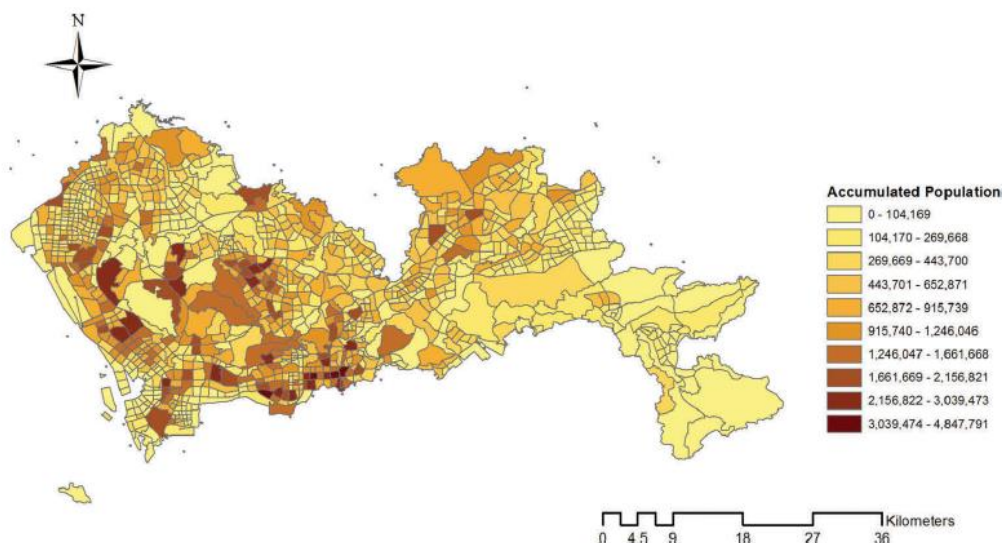
- EVP – těch je hodně a málo četně se tam jezdí
- Průměrný počet OVP je 5 (CDR-67) (typicky hospoda, rodiče, sport),
- MVP jen 2 (doma a práce).



- zastoupení cest mezi cíli:
 - Milán (CDR-17): 61% cest mezi MVP (MVP-MVP) a 17% pro OVP-OVP.
 - Region (CDR-67): 46 % cest mezi MVP, 9% mezi OVP a 9% mezi OVP a MVP.
- 5000 simulací individuálních cest pomocí Markovových řetězců (začíná z nejoblíbenějšího místa, zvolí se kategorie dalšího cíle, výběr konkrétního POI podle RR, spočítá se doba cesty jednoduše vynásobením vzdálenosti průměrnou rychlostí).
- porovnání doby setrvání v buňce podle modelu a podle reality pomocí Hellingerovy vzdálenosti mezi 2 pravděpodobnostními distribucemi



- Yang (2017): oblast města Shenzen (Čína), 15 mil na 2000 km² (Praha 500 km²). Počet uživatelů mobilu v pracovní den jako proxy místní vitality, po půl hodině.
- Aktivní zjišťování SIM
- Pouze 1 den (pátek) v 2012, 1 z hlavních mobilních operátorů – přesto záznam asi 12 mil. uživatelů (80% populace). Celkově asi 6000 BTS.
- terénní výzkum - Thiessenovy polygony pro BTS neodpovídají realitě a výrazně nadhodnocují počty! Např. v 1 místě skutečná populace pod 100, z výpočtu několik tisíc.
- územní jednotky TAZ (traffic analysis zones) (počty obyvatel a údaje o zaměstnanosti). Celkem 1112 TAZ s průměrnou plochou 1.8 km².
- identifikovali BTS v každé TAZ a sečetli všechny záznamy z těchto BTS jako 24h populaci
- Adjustace na počet obyvatel, velikost a zaměstnanost v TAZ (hustota obyvatel, hustota zaměstnanců, poměr zaměstnanci k obyvatelům)
- Měření diverzity území TAZ pomocí Hill čísel (indikátory bohatosti, neuspořádanosti a koncentrace).



Děkuji za Vaši pozornost



jiri.horak@vsb.cz

