

# Dokumentace k projektu UPA

## Zvolené téma

COVID-19 (varianta 04) – dr. Rychlý

## Řešitelé

- Filip Jeřábek (xjerab24)
- Daniel Konečný (xkonec75)
- Tomáš Ryšavý (xrysav27)

## Zvolené dotazy a formulace vlastního dotazu

### Dotaz varianty A

V grafech zobrazte tempo změny počtů aktuálně nemocných (absolutní i procentuální přírůstek pozitivních případů a klouzavý průměr různých délek v různých časech).

### Dotaz varianty B

Určete vliv počtu nemocných a jeho změny v čase na sousední okresy (aneb zjistěte jak se šíří nákaza přes hranice okresů).

### Vlastní dotaz (C)

Lokalizujte ohniska přírůstku nakažených nemocí COVID-19. Dotaz u každého okresu porovná jeho denní přírůstek s denními přírůstky všech sousedních okresů. Pokud je jeho přírůstek ze všech nejvyšší, je daný okres označený jako ohnisko.

## Stručná charakteristika zvolené datové sady

Zdrojem dat jsou otevřené datové sady publikované Ministerstvem zdravotnictví České republiky týkající se onemocnění COVID-19 dostupné na adrese: <https://onemocneni-aktualne.mzcr.cz/api/v2/covid-19>. Používaná data jsou z datové sady s názvem COVID-19: Přehled epidemiologické situace dle hlášení krajských hygienických stanic podle okresu a zpracovaná jsou ve formátu JSON, který je dostupný na této adrese: <https://onemocneni-aktualne.mzcr.cz/api/v2/covid-19/kraj-okres-nakazeni-vyleceni-umrti.json>. Datová sada obsahuje denní informace o přírůstku nakažených, vyléčených a zemřelých osob v jednotlivých krajích a okresech. Je aktualizovaná s týdenním zpožděním z důvodu validace informací krajskými hygienickými stanicemi.

Data ve formátu JSON jsou stažena z internetové adresy uvedené výše. Kořenový JSON objekt obsahuje informaci o čase vytvoření této datové sady, dále zdroj, ze kterého data pocházejí, a poté pole samotných dat o přírůstcích. Jedna položka tohoto pole odpovídá informacím z jednoho dne v jednom okrese a navíc je uveden i kraj. Dále obsahuje kumulativní počet nakažených, vyléčených a zemřelých.

```
{
  "modified": "2020-10-21T01:01:32+02:00",
  "source": "https://onemocneni-aktualne.mzcr.cz/",
  "data": [
    {
      "datum": "2020-03-01",
      "kraj_nuts_kod": "CZ010",
      "okres_lau_kod": "CZ0100",
      "kumulativni_pocet_nakazenych": 2,
      "kumulativni_pocet_vylecenych": 0,
      "kumulativni_pocet_umrti": 0
    },
    {
      "datum": "2020-03-01",
      "kraj_nuts_kod": "CZ020",
      "okres_lau_kod": "CZ020A",
      "kumulativni_pocet_nakazenych": 0,
      "kumulativni_pocet_vylecenych": 0,
      "kumulativni_pocet_umrti": 0
    },
    ...
  ]
}
```

Dále se využívá námi vytvořená datová sada, která obsahuje informace o tom, který okres sousedí se kterým. A datová sada přiřazující názvu okresu LAU kód, která pochází z webu Ministerstva zemědělství České republiky. Datová sada se nachází zde:

<https://eagri.cz/ssl/nosso-app/DataKeStazeni/Okresy?page=1&sortBy=NAZEV~NAZEV&pageSize=1000&collapsed=False>

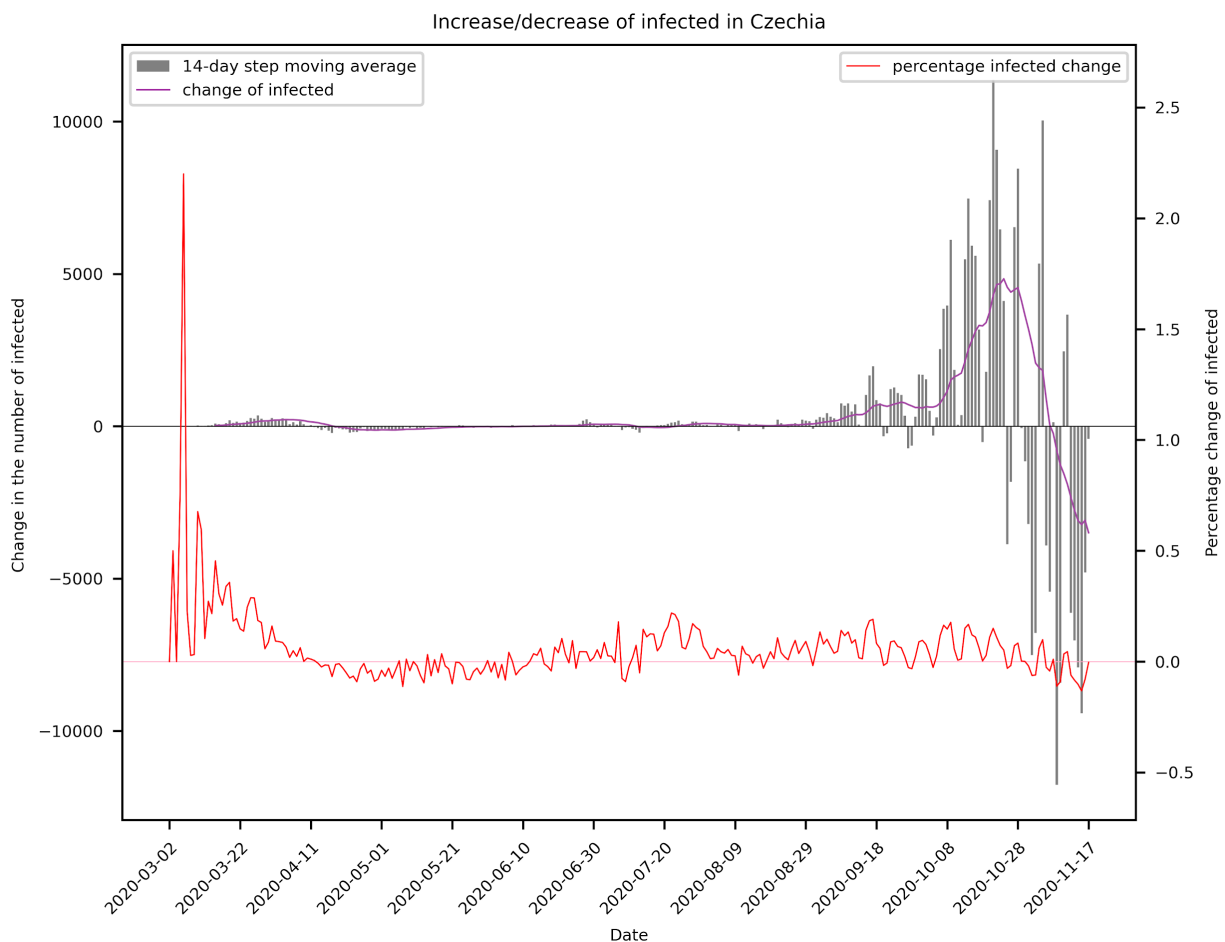
## Zvolený způsob uložení surových dat

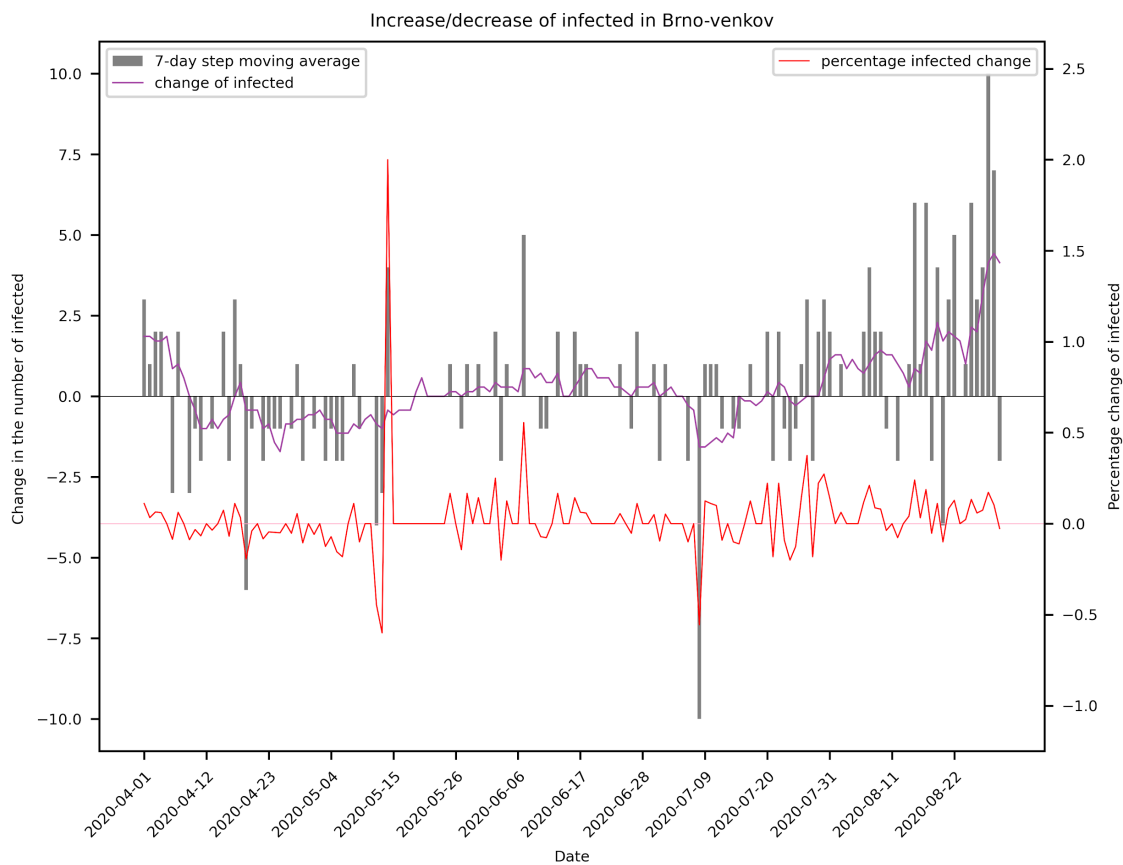
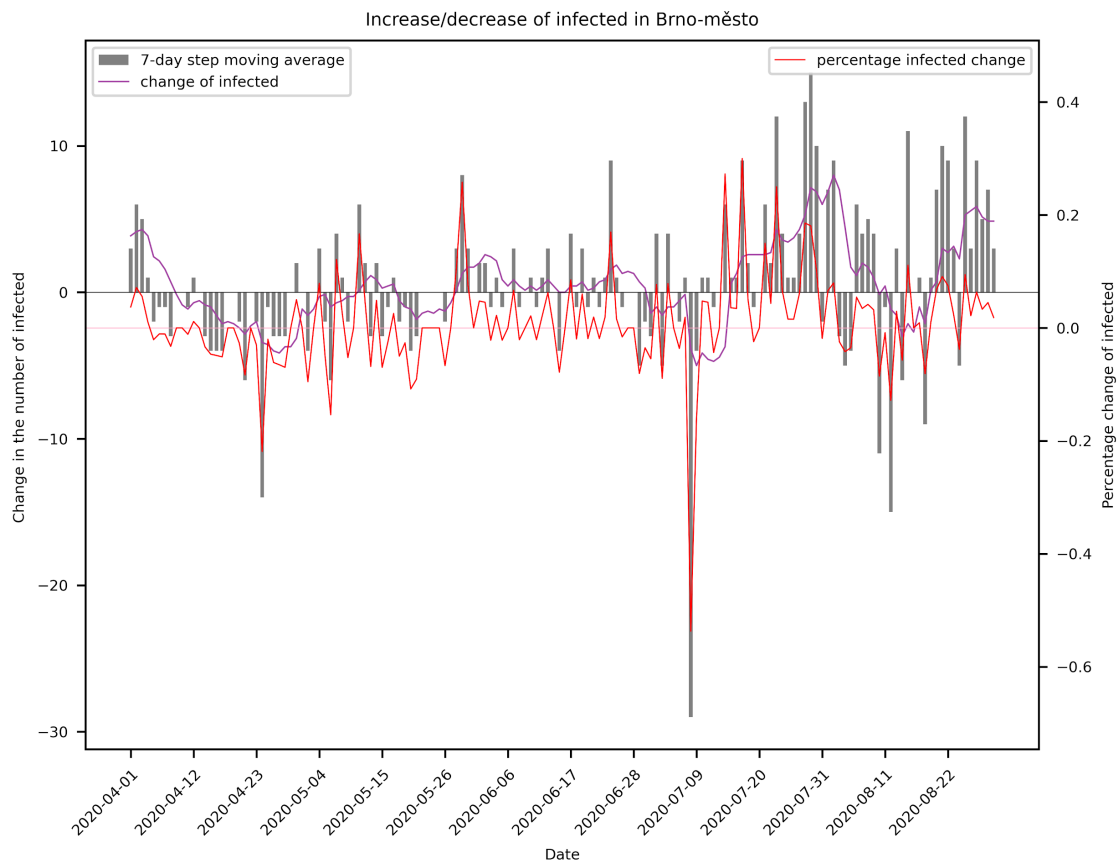
Data budou uložena v NoSQL databázi Neo4j. Každý denní záznam v jednom okrese reprezentuje jeden uzel. Sousedící okresy jsou propojeny patřičnou vazbou. Dále jsou postupně propojeny vazbou jednotlivé denní záznamy náležející jednomu okresu. V každém uzlu je uveden kumulativní počet nakažených, vyléčených a zemřelých jako vlastnost tohoto uzlu.

# Řešení a výsledky dotazů

## Dotaz varianty A

Je implementován dotaz, který získá absolutní i procentuální přírůstek v daném okrese v dané datum a tato data jsou poté importována do vlastní tabulky relační databáze (SQLite). Přírůstek je třeba počítat z kumulativního počtu nakažených, vyléčených a zemřelých. Dále je implementován druhý dotaz, který vypočítá klouzavý průměr přírůstků a uloží je do další tabulky relační databáze. Tyto průměry jsou vypočítané pro 4 možné délky klouzavého průměru (3, 7, 14 a 28 dní), tedy je možné pracovat pouze s nimi. Pro jiné intervaly lze pouze jednoduše změnit délku průměru ve skriptu `nosql_to_relation.py`.





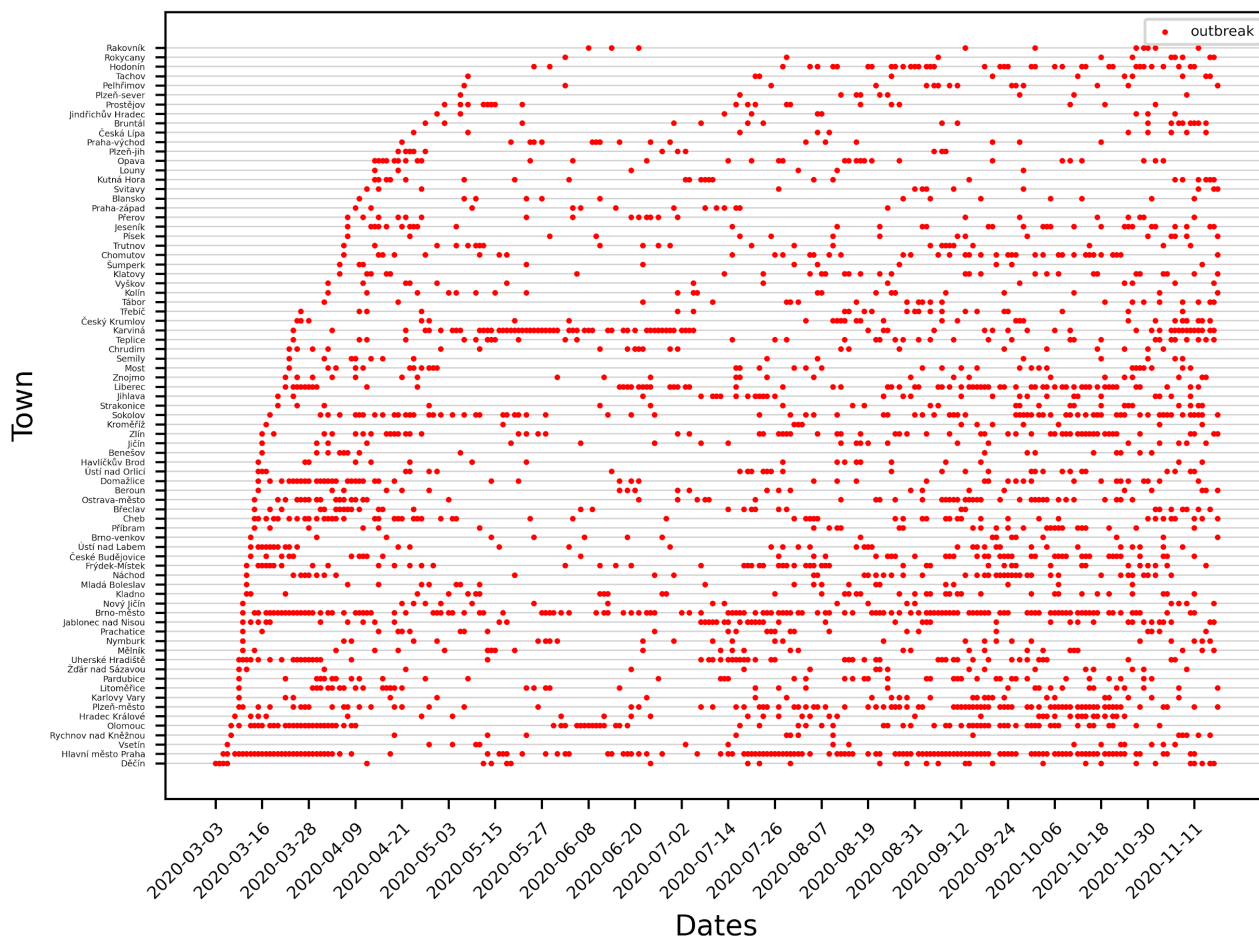
## Dotaz varianty B

Dotaz B není implementován z důvodu obtížnosti a nedostatku znalostí z oboru epidemiologie a matematického modelování. Z pohledu získávání dat (sousedních uzlů a přírůstků) není příliš mnoho principiálních rozdílů mezi dotazem B a vlastním dotazem (C). Byl by tedy využit stejný mechanismus získání sousedů pomocí přechodů mezi okresy a dny. Problém je však ve vyhodnocení, zda některý okres má doopravdy vliv na zvýšení přírůstku nakažených. Je třeba testovat, kterým okresem je konkrétní okres ovlivněný, a jestli takto ovlivňuje i jeho další sousední okresy (tím se potvrdí, že reálně ovlivňuje sousedy). Dále je potřeba určit, jaká hodnota už reálně je ovlivňující a jaká je pouze normální zvýšený přírůstek. A nakonec je potřeba vzít v potaz, že ovlivnění se ukáže až s několikadenním zpožděním, než se nakažení lidé nechají otestovat.

## Vlastní dotaz (C)

Dotaz C se provádí nad jedním konkrétním uzlem v databázi, který představuje okres v konkrétním dnu. Získáním všech sousedních okresů a uzlů předchozích dnů všech těchto okresů jsme schopni získat maximální přírůstek nakažených v okolních okresech a porovnat ho s přírůstkem testovaného okresu. Pokud je hodnota testovaného uzlu větší než jeho sousedů, okres splňuje podmínky ohniska a je vrácen jako výsledek dotazu. Dále je vrácena hodnota přírůstku a rozdíl od druhého nejvyšší přírůstku v sousedních okresech.

When towns were outbreaks in Czechia



When towns were outbreaks in selected towns

