1. Dataset
   1. Covid-19 Search Trend Symptoms Dataset

The Covid-19 Search Trend Symptoms Dataset was sourced from a public repository maintained by Google Research. It is an aggregated, anonymized dataset which shows trends in search queries for symptoms related to Covid-19.

The data reflects the volume of Google searches for a broad set of health symptoms and is organized by geographic region. The resulting dataset is a daily timeseries for each state in the United States of America (USA) showing the relative frequency for searches for each symptom. The following table lists the relevant Covid-19 symptoms found in the dataset.

|  |  |
| --- | --- |
| Symptoms | Synonyms |
| Ageusia | loss of taste |
| Anosmia | loss of smell |
| Arthralgia | joint ache |
| Common Cold | viral rhinitis |
| Cough | tussis |
| Diarrhea | dysentery |
| Fatigue | exhaustion |
| Fever | pyrexia |
| Headache | cephalgia |
| Myalgia | muscle ache |
| Shortness of Breath | breathing difficulty |
| Sore Throat | pharyngitis |
| Nausea | sensation of being about to vomit |
| Vomiting | puking |

Table 1: Relevant Covid-19 Symptoms according to Centre of Disease Control and Prevention (CDC)

Let P be the considered period which is defined as Let R be the states in the USA (including Washington DC) of size 51. R is formally defined as . For the purpose of this research, we will call these states, ‘regions’. Table 2 below lists the 51 states and their formal abbreviations. Let S be the relevant Covid-19 symptoms which is defined as . Therefore, we define the Covid-19 Search Trend Symptoms Dataset as whereby P, R, S is of size p, 51 and 14 respectively.

|  |  |
| --- | --- |
| State | Abbreviation |
| Alaska | AK |
| Alabama | AL |
| Arkansas | AR |
| Arizona | AZ |
| California | CA |
| Colorado | CO |
| Connecticut | CT |
| Mississippi | MS |
| Montana | MT |
| North Carolina | NC |
| North Dakota | ND |
| Nebraska | NE |
| New Hampshire | NH |
| New Jersey | NJ |
| New Mexico | NM |
| Indiana | DC |
| Delaware | DE |
| Florida | FL |
| Georgia | GA |
| Hawaii | HI |
| Iowa | IA |
| Idaho | ID |
| Illinois | IL |
| Indiana | IN |
| Kansas | KS |
| Kentucky | KY |
| Louisiana | LA |
| Massachusetts | MA |
| Maryland | MD |
| Maine | ME |
| Michigan | MI |
| Minnesota | MN |
| Missouri | MO |
| Nevada | NV |
| New York | NY |
| Ohio | OH |
| Oklahoma | OK |
| Oregon | OR |
| Pennsylvania | PA |
| Rhode Island | RI |
| South Carolina | SC |
| South Dakota | SD |
| Tennessee | TN |
| Texas | TX |
| Utah | UT |
| Virginia | VA |
| Vermont | VT |
| Washington | WA |
| Wisconsin | WI |
| West Virginia | WV |
| Wyoming | WY |

Table 2: Elements of R

* 1. Covid-19 Cases Dataset

The Covid-19 Cases Dataset was sourced from The Covid Tracking Project, a volunteer organization launched from The Atlantic. The dataset is a daily timeseries, for each region *r*, of the daily increase in confirmed Covid-19 cases. Let P be the considered period which is defined as and let R be the states in the USA, formally defined as and |R|=51. Thus, we define the Covid-19 Cases Dataset as .

1. Method
   1. Linear regression model for symptom rate prediction

For a target region and a single symptom , we use day *d* symptom rate to predict day *d+1* symptom rate using a linear regression model which is defined in the following subsections.

* + 1. Obtaining Weights

For each symptom and each day , the symptom rate for target region is predicted using the symptom rate from another 5 control regions which have the highest coefficient of determination R2 and are not bordering the target region, .

Subsection 1.1

Subsection 1.2

**Experiment 2**

Initial Algorithm

L is calculated over 330 days using 330 days of Actual Case Counts

1. L\_hat is calculated over 300 days

L\_hat is then compared to 300 days window of L (based on lag)

New Algorithm

1. 300 day window of Actual Case Counts is calculated
2. L is calculated over 300 day window of Actual Case Counts (based on lag)
3. L\_hat is then compared to L