This document contains information and guidance about the data you will be using for your models. All the data is publicly available. This is based off the data I have been using for my models and although I have the “clean” data, you should have a go at cleaning the raw data and processing covariates yourself. I have guidance in this document and some scripts to help with the more complicated side of things. Any questions, please let me know!

**Norovirus data:**

* norovirus\_stw\_raw.csv and site\_info.csv have been downloaded from the CEFAS portal: <https://data.cefas.co.uk/view/21753>
* The combined NoV concentration will be the outcome of the model. Don’t worry if the column descriptions file are confusing, you won’t need most of the variables.
* Site\_info data contains the location and population of the sewage treatment works (STW) catchments
* It is important to have a think about how you might clean the data, dealing with missingness, duplicates etc. There is no right answer, we can discuss it when we meet 😊

**Shapefiles:**

* A shapefile for the STW catchment is located in the STW folder, note that for the south west of England, the catchments had to be approximated (see the Li et al paper for more details on this)
* For the model prediction we will use LSOAs as our prediction grid, which are very fine spatial resolution. You can find the shapefiles for the 2021 boundaries and population weighted centroids from the GOV geo portal: <https://geoportal.statistics.gov.uk/>. Make sure it’s the 2021 boundaries as this is the year most of the NoV data was sampled.
* The centroids will be the grid and the LSOA boundaries are for mapping
* Note: if we have enough time in the project we will aggregate the predictions from LSOA to local authority districts, as this is more informative from a policy perspective

**Covariates:**

These covariates have been selected for the NoV model based on literature. If you want to use any other data, you are welcome to do so!

**Ethnicity:**

Source: 2021 Census

Proportion of non-white ethnicities by LSOA

**Index of multiple deprivation:**

Source: 2021 Census

Index of multiple deprivation (IMD). LSOA are ranked from most deprived (1) to least. This metric is generally based off income, employment, education, health, crime, barriers to housing and services, and the living environment.

**School and carehome locations:**Source: UK government

Location of schools and carehomes in England. These data can be processed as densities (number per 100 000) people. See the code for guidance.

**Land use:**

Source: CORINE Land Cover <https://land.copernicus.eu/en/products/corine-land-cover>

Land use categories per 100m x 100m in England. More information can be found from the data source. I have attached the grib file in the “Land use” folder. Like Li et al, you can process this so you get the proportion of each land cover category for the STW catchment and LSOAs. I have been using urbanicity and agriculture use in my models. We hypothesised that agricultural areas may confound the norovirus concentration as ammonia in fertilisers may affect the flow normalisation process. Check the code for what numbers correspond to what categories, its up to you which land use categories you want to include in the model but I wouldn’t recommend doing 2-3.

**Temperature and rain:**

Source: ERA5 climate data store <https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels?tab=overview>

This can be a pain to download so I have given you the raw data. This involves hourly rain (total precipitation) and temperature (2m temperature). The spatial resolution is a regular lat-lon grid of 0.25 degrees. I have given you the code to get the daily mean and extract this from the STW catchments and LSOAs (using terra::extract).

**Population mobility data:**

Source: Google COVID-19 population mobility reports <https://www.google.com/covid19/mobility/>

Population mobility data can be sourced from the google COVID-19 mobility reports in the form of percentage change in travel to work places from baseline (before COVID-19 pandemic). More information can be found on the website.

**Lockdown stage:**

Source: <https://acrobat.adobe.com/id/urn:aaid:sc:EU:e653671b-5343-4449-a0f1-588003a08387>   
<https://acrobat.adobe.com/id/urn:aaid:sc:EU:7a27ca31-685b-4a93-bdf8-36285873ce74>   
  
Taken from various COVID-19 lockdown timelines. I’ve attached the code for the different stages so you don’t have to hunt through the documents.

**Processing data:**

* For the covariates available at LSOA, you can align these to the STW catchments using a population weighted average of the intersecting LSOAs with the STW catchments
* It isn’t as straight forward for the land use (urbanicity and agriculture) and meteorological (temperature and rain) covariates. I have attached the code to extract these (see “Code” folder). See Li et al paper for how this is processed.
* For covariates available daily, have a think about how you will align this i.e. lags, rolling averages etc.