Starting Guide

```
<img src="../../docs/images/covid-droplet.svg" alt="Logo" style="width:200px">
```

Getting Started

Overview

This package is intended to make the process of analyzing epidemiological wastewater data easier and more insightful. It's intended primarily for researchers and epidemiologists but could potentially be used by anyone with an interest in the topic who has a working knowledge of R programming, epidemiology, and statistics.

Requirements: - Knowledge of R programming - Familiarity with epidemiology - A base understanding of statistics

This package provides a set of core utilities for preparing your data, analyzing your data, and visualizing the results.

Features: - Outlier detection - Various data smoothing techniques - Normalization techniques - Time series analysis - Wastewater / case offset analysis - Visualization / graphing tools - Sample data sets courtesy of the Wisconsin Department of Health Services

Note that the package also includes a set of example data. This was done in order to provide a set of real-world instructive examples which make the code easier to understand and apply to your own data sets.

What is Wastewater Based Epidemiology?

Epidemiology is the process of investigating and monitoring the prevalence of disease agents within a population or an environment. This has been most commonly performed by collecting case data from hospitals and other public health agencies. Wastewater based epidemiology takes a different approach by looking for disease agents in the population's wastewater as it is collected in sewage treatment plants.

Problems with Traditional Case Based Epidemiology

During the Covid-19 pandemic, many problems associated with the traditional case based epidemiology approach became evident.

- Variations in testing
- Variations in reporting
- Privacy concerns associated with collection on individual health data
- Disease can take time before onset and presentation in a doctor's office so the technique is not very timely.

The Process of Wastewater Based Epidemiology

To address these problems, attention has shifted to an alternate / complementary approach - wastewater based epidemiology. Wastewater epidemiology is conducted using the following process:

- Wastewater samples are periodically collected from sewersheds
- Samples are sent to state laboratories for analysis
- Analysis results are communicated to the state health services agencies, where they are compared with reported case rates

```
<img src="../../docs/images/getting-started/wastewater-process.png" alt="The Wastewater Epidemiology Pro
<div>
    <label>The Wastewater Epidemiology Process</label>
</div>
```

Benefits of Wastewater Based Epidemiology

The advantages of wastewater based epidemiology compared with case based epidemiology are as follows:

- Can provide more timely, almost “real-time” information
- Is well suited for an epidemiological early warning system
- Maintains the anonymity of individuals
- Does not rely upon voluntary testing by individuals so it has the potential to be more consistent.

```
<img src="../../docs/images/getting-started/wastewater-offset.png" alt="Onset of Symptoms and Wastewate
<div>
    <label>Onset of Symptoms and Wastewater Detection</label>
</div>
```

Challenges Associated with Wastewater Based Epidemiology

Despite the very attractive characteristics and promising results of wastewater based testing, there are also a number of potential challenges associated with this approach which can make it difficult to implement:

- Data is inherently noisy
- There are often significant sampling differences between communities (once per day verses once per week, for example)
- There are often differences in methodology (qpcr verses dpcr etc.)
- There are many cofactors related to wastewater collection and testing which can make interpretation of results difficult.

```
<img src="../../docs/images/getting-started/sample-frequency.png" alt="Differences in Sampling Frequency
<div>
    <label>Differences in Sampling Frequency</label>
</div>
```

The Role of This Software

Because of these various complicating factors and difficulty in performing wastewater based analysis and interpreting results, software such as this can serve as a valuable aid in making the analysis and interpretation of this data easier and more reliable.

Loading and Viewing Data

The data in this package is a combination of data provided to us from the following sources: - Wisconsin Department of Health Services (DHS) - Wisconsin State Lab of Hygiene (SLH) - Open-source data

All data can be found in the /data directory as .RData objects. Alternatively, when our package is installed, these data sets can be loaded by using the command:

```
data(<name here>, package = "Covid19Wastewater")
```

where <name here> is replaced with one of the following:

Data List

- Aux_info_data Extra data that can be merged with WasteWater_data
- Case_data Case information for all of Wisconsin from 2020-01-22 to 2022-12-08
- Covariants_data Statewide variant proportions
- Example_data A merged and shortened version of Case_data and WasteWater_data from 3 sites
- HFGCase_data High-frequency data from 6 weeks involving ten sites
- HFGWaste_data High-frequency data from 6 weeks involving ten sites
- InterceptorCase_data Madison specific data
- Pop_data Population data along with region, county, and lab submitter
- WasteWater_data Wastewater epidemiological data from across Wisconsin, can be merged with Aux_info_data

Here is the key to all the column names in the data: https://github.com/UW-Madison-DSI/Covid19Wastewater/blob/main/docs/data/data_columns_discription.md

Examples

Viewing Gene Markers Over Time

The prevalence of covid is determined using the genome markers called “N1” and “N2”. A simple starting point is to load in the data and then graph N1 or N2 over time.

```r
data("WasteWater_data", package = "Covid19Wastewater")

WasteWater_data %>% ggplot(aes(x=date,y=N1)) +
  geom_point()
```

```
<img src="../../docs/images/getting-started/n1-n2-levels.png" alt="The Levels of Covid Makers N1 and N2
<div>
    <label>The Levels of Covid Markers N1 and N2 Over Time</label>
</div>
```

Adding Color

With a few extra lines of code, we can add some color coding in order to display N1 and N2 in a more visually appealing way.

```
data("WasteWater_data", package = "Covid19Wastewater")

WasteWater_data %>% ggplot() +
  geom_point(aes(x=date,y=N2+1, color = "N2")) + #plus 1 to have a nice log
  geom_point(aes(x=date,y=N1+1, color = "N1")) +
  scale_y_log10() +
  ylab("N1 and N2")
```

```
<img src="../../docs/images/getting-started/n1-n2-levels-colored.png" alt="The Levels of Covid Makers N
<div>
    <label>The Levels of Covid Markers N1 and N2 Over Time</label>
</div>
```

Merging Datasets

Below, we show a set of examples of merging different datasets together.

1. Merging Wastewater and Case Data

When merging wastewater and case data, it is best to merge by site and data to identify each entry uniquely.

```
data("WasteWater_data", package = "Covid19Wastewater")
data("Case_data", package = "Covid19Wastewater")

WasteAndCaseMerged_data <- merge(Case_data,WasteWater_data, by = c("site","date"))
head(WasteAndCaseMerged_data)
```

```
<img src="../../docs/images/getting-started/wastewater-case-data.png" alt="Wastewater and Case Data" sty
<div>
    <label>Wastewater and Case Data</label>
</div>
```

2. Merging High Frequency Wastewater and Case Data

We include high frequency datasets for both waterwater and case data which can be merged as follows:

```
data("HFGWaste_data", package = "Covid19Wastewater")
data("HFGCase_data", package = "Covid19Wastewater")

HFGWasteAndCaseMerged_data <- merge(HFGCase_data,HFGWaste_data, by = c("site","date"))
head(HFGWasteAndCaseMerged_data)
```

```
<img src="../../docs/images/getting-started/hfg-wastewater-case-data.png" alt="High Frequency Wastewate
<div>
    <label>High Frequency Wastewater and Case Data</label>
</div>
```

3. Merging Wastewater and Aux Data

When merging the auxiliary information, it can only be done with sample_id (Aux_info_data can only be merged with WateWater_data)

```r
data("WasteWater_data", package = "Covid19Wastewater")
data("Aux_info_data", package = "Covid19Wastewater")

WastewaterAndAuxInfo_data <- merge(WasteWater_data,Aux_info_data, by = "sample_id")
head(WastewaterAndAuxInfo_data)
```

```
<img src="../../docs/images/getting-started/wastewater-aux-data.png" alt="Wastewater and Aux Data" style
<div>
    <label>Wastewater and Aux Data</label>
</div>
```

4. Merging Wastewater and Population Data

Population data can be merged with any dataframe that contains site data.

```r
data("WasteWater_data", package = "Covid19Wastewater")
data("Pop_data", package = "Covid19Wastewater")
data("Case_data", package = "Covid19Wastewater")

WastewaterAndPop_data <- merge(WasteWater_data,Pop_data, by = "site")
head(WastewaterAndPop_data)

CaseAndPop_data <- merge(Case_data,Pop_data, by = "site")
head(CaseAndPop_data)
```

```
<img src="../../docs/images/getting-started/wastewater-pop-data.png" alt="Wastewater and Population Data
<div>
    <label>Wastewater and Population Data</label>
</div>
```

4. Merging Wastewater and Confirmed Cases Data

With the data now merged, we can perform many more analyses. In the analysis below, we show the number of confirmed cases.

```r
HFGWasteAndCaseMerged_data %>% ggplot() +
  geom_point(aes(x=log(N1+1),y=log(ConfirmedCases+1),color="N1")) +
  geom_point(aes(x=log(N2+1),y=log(ConfirmedCases+1),color="N2")) +
  facet_wrap("site")
```

```
<img src="../../docs/images/getting-started/wastewater-confirmed-cases.png" alt="Wastewater and Confirme
<div>
    <label>Wastewater and Confirmed Cases</label>
</div>
```

4. Normalizing by Population

Below, we display the number of confirmed cases normalized by population as a function of time.

```
CaseAndPop_data %>% filter(site == "Madison") %>% ggplot(aes(x=date,y=(conf_case/pop)))+
  geom_point()
```

```
<img src="../../docs/images/getting-started/confirmed-cases-per-person.png" alt="Wastewater and Confirme
<div>
    <label>Wastewater and Confirmed Cases / Population</label>
</div>
```

Tips

Always make sure that when merging, the "by =" should always be able to identify the information you are merging uniquely. (i.e. don't merge waste and case data by date alone)

Data Preparation

The data preparation takes two main forms: - Outlier detection and Removal - Smoothing methods

Smoothing methods

There are three smoothing methods available to get a more stable Wastewater measurement - loessSmoothMod - expSmoothMod - sgolaySmoothMod Each one can generate a consistent signal from weekly data. A comprehensive Guide on the methods is available here

Outlier detection

There are two main ways to detect outliers in this package. - Deviance from the trend - Unusual spikes from adjacent values ### Deviance from the trend This process has two steps. First you need a trend. This can normally be done with the smoothing in the previous section. Then the trend can be used to find points sufficiently greater than it. This is normally set to 2.5 standard deviations.

```
data("WasteWater_data", package = "Covid19Wastewater")
WasteWater_data  <- filter(WasteWater_data, site == "Janesville")
WasteWater_data <- mutate(WasteWater_data, N1 = log(N1 + 1))
WasteWater_data <- loessSmoothMod(WasteWater_data , "N1", "N1_loess")
WasteWater_data <- Flag_From_Trend(WasteWater_data,  N1, N1_loess)

WasteWater_data %>%
  ggplot(aes(x = date))+
  geom_point(aes(y = N1, color = flagged_outlier))+
geom_line(aes(y = N1_loess, color = "N1 Loess"))+
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
  labs(y = "Covid-19 Gene Concentration",
       x = "Date",
       color = "Flagged Outlier"
       )
```

Unusual spikes from adjacent values

Data Analysis

Time Series Analysis

Shedding is an ongoing process from the first day of infection to days or even weeks after symptoms subside. Thus it is hard to know exactly how many individuals in the community are infected at any given time just using wastewater data. Since we have mostly reliable case data, if we can find the offset that best correlates with the 2 data sets, we can work backward from only wastewater data in the future.

```
<img src="../../docs/images/getting-started/time-series-analysis.png" alt="Heatmap of waste to case cor
<div>
    <label>Heatmap of waste to case correlation</label>
</div>
```

This analysis was done by finding the R-squared correlation between the wastewater data of the current day Z and the combined case data from past Y many days and X number of future days. Thus we can find which moving window of days best represents the wastewater data for the next analysis.

```
<img src="../../docs/images/getting-started/offset-analysis.png" alt="Offset analysis" style="width:75%
<div>
    <label>Offset analysis</label>
</div>
```

Using the window of case data days that best correlates to the wastewater data, we can find the offset that best corresponds to the time between shedding at its peak and when the individual got tested.

Conclusion

We hope that you have had a successful and enjoyable experience using this software package. If you would like to share your results and/or feedback with the package authors, contact information is listed below:

- Marlin Lee - (mailto:mrlee6@wisc.edu)
- Kyllan Wunder - (mailto:kwunder@wisc.edu)
- Abe Megahed - (mailto:amegahed@wisc.edu) You may also submit comments, feedback, feature requests, and bug reports through the GitHub repository at: https://github.com/UW-Madison-DSI/Covid19Wastewater

Acknowledgements

This package was made possible through support from the University of Wisconsin Data Science Institute, in collaboration with the Wisconsin Department of Health Services (DHS), and the State Lab of Hygiene (SLH).