/\*WVAL Analysis Code for Covid, RSV, Flu A and Flu B\*/

/\*Created on 07/19/2024\*/

/\*Created by John C Webster - Adapted from Isabel Damazo\*/

/\*Last Updated 01/15/2025\*/

\*---------------------------------------------------------------------------;

/\*SAS ANALYSIS FOLDER: this is where we store the code itself, and any other subfolders created in the project. \*/

/\*Raw Data Folder: This is where we pull in the raw WW data that comes from our partners each week. \*/

\*---------------------------------------------------------------------------;

/\* ESTABLISH A LIBRARY FOR WW FILES \*/

/\* Change path to your computer directory if necessary \*/

%let path = *your\_computer\_path*\SAS Analysis;

%let path2 = *your\_computer\_path*\Raw Data;

\*---------------------------------------------------------------------------;

/\*Create a Library for Disease Data\*/

libname Disease "&path";

**run**;

\*---------------------------------------------------------------------------;

/\*START OF THE MACRO PROGRAM\*/

options spool mlogic symbolgen; \*these options here are optional, they just help when checking errors in the macro if they occur;

**%Macro** Disease(libname=,ImportSheet=,Disease=);

\*---------------------------------------------------------------------------;

/\* MACRO STEP 1 - IMPORT FILES \*/

/\* Import Appropriate Disease Raw Data - Ours is an Excel Workbook so we use the libname xlsx import engine to be able to work out of different sheets. \*/

options validvarname=v7;

proc import datafile="&path2\&ImportSheet..xlsx" out=All\_Regions

dbms=xlsx replace; sheet="All Regions";

run;

/\* Import your Historical Flow Data, ours is kept in a separate excel sheet which is why we do this, but if you have all the flow data in your raw collection files then you can skip any of the flow data wrangling in subsequent steps. \*/

options validvarname=v7;

proc import datafile="&path2\flow\_mgd\_all.xlsx" out=flowdata

dbms=xlsx replace; sheet="Flow";

run;

/\* Some of our flow data brought the site in with “WWTP” attached to it, so I needed to drop that for merging purposes later. See example below. \*/

data flowdata;

set flowdata;

length name $**20**;

if wwtp\_name = "Ada WWTP" then name="Ada";

else if wwtp\_name = "Altus WWTP" then name="Altus";

drop wwtp\_name;

format sample\_collect\_date mmddyy10.;

run;

/\*Keep only relevant variables for analysis purposes. Here I keep only the site name, ID, virus per liter reading, and the collection date,\*/

data all\_regions;

set all\_regions(keep=ID--virus\_l);

run;

data all\_regions;

set all\_regions;

if virus\_l="no data" then virus\_l=**.**;

run;

data all\_regions;

set all\_regions;

format Sample\_Date mmddyy10.;

run;

/\*Here I added in our state’s HHS Regions to the Data. See example below. I omitted some of the regions here so you didn’t have to try and look at all 8 of them. Just be sure to add in however many regions you all have.\*/

data All\_Regions;

set All\_Regions;

log\_Virus=log(virus\_l);

if substr(Name,**1**,**3**)='OKC' or Name in ('Midwest City','Edmond') then HHS\_Region=**8**;

else if substr(Name,**1**,**5**)='Tulsa' or Name='Bixby' then HHS\_Region=**7**;

else if name in ('Yukon','Norman','Purcell') then HHS\_Region=**6**;

run;

/\*Add in County to the Data. This is basically the same process as the HHS region. Just create a new “County” variable and assign the site to its appropriate county it resides in. Again, I included a subset of examples below.\*/

data All\_Regions;

set All\_Regions;

length County $**20**;

if substr(Name,**1**,**3**)='OKC' or Name in ('Midwest City','Edmond') then County="Oklahoma";

else if substr(Name,**1**,**5**)='Tulsa' or Name='Bixby' then County="Tulsa";

else if name in ('Purcell') then County="McClain";

run;

/\* Merge Flow Data to All\_Regions – again, if your flow data is already in your raw files you can skip this step. \*/

proc sql;

create table all\_regions1 as

select a.\*, b.flow\_rate

from all\_regions a

left join

flowdata b

on a.Sample\_date=b.sample\_collect\_date and a.Name=b.name;

quit;

proc sort data=All\_regions1 noduprecs;

by sample\_date;

run;

/\*Omit sites we no longer sample – we had some contracted sites in early 2021-2022 that we no longer sample so I am omitting them for when we do an analysis on a historical subset of the data to derive values needed for the WVAL calculation. Here I use the NOT IN statement, so any values not included in the list will be dropped. \*/

data all\_regions1;

set all\_regions1;

if name not in("Ada", "Anadarko", "Bartlesville", "Durant", "Enid", "Midwest City", "Muskogee", "Norman", "OKC Deer Creek", "OKC North Canadian", "Seminole", "Stillwater", "Tahlequah", "Tulsa Haikey Creek", "Tulsa North", "Tulsa South", "Woodward", "Yukon") then delete;

run;

/\*Flow Normalize Virus/L concentrations – Here I define a function much like you would do in other languages like R or Python. To do this you specify some keyword parameters you want to pass into the function (Virus\_L, flow\_rate, Population). The function does the appropriate math then returns the value you ask for. I then call the defined function in the next step.\*/

proc fcmp outlib=work.normalize.flow;

function normalizer(Virus\_L, flow\_rate, Population);

flow\_normal = flow\_rate \* **3.7854** \* **1000000** ;

virus\_flow = Virus\_L \* flow\_normal;

Virus\_L\_Normalized = virus\_flow/Population;

return(Virus\_L\_Normalized);

endsub;

options cmplib=work.normalize;

data all\_regions1;

set all\_regions1;

Virus\_L\_Normalized = normalizer(Virus\_L, flow\_rate, Population);

run;

\*---------------------------------------------------------------------------;

/\* MACRO STEP 2 - SEPERATE DATA INTO CURRENT AND HISTORICAL TABLES \*/

/\*Divide the data into a current table and a historical table – I do this because I will be deriving some summary statistics like the historical average, historical standard deviation, and the 10th percentile of historical averages to use in the WVAL calculation\*/

data &Disease.\_Current;

set All\_Regions1;

run;

data &Disease.\_Historical;

set All\_Regions1;

where Sample\_Date lt **"&sysdate"d** - **7** ; /\*This tells the historical table to include everything except the current week’s data \*/

run;

\*---------------------------------------------------------------------------;

/\* MACRO STEP 3 - USE MACRO PROGRAMMING TO CREATE SITE SPECIFIC TABLES\*/

/\*Obtain a list of sample sites stored into a macro variable.\*/

proc sql noprint;

select distinct compress(name) into :Site1-

from &Disease.\_Current;

quit;

/\*Obtain a count of the sample sites for do-loop processing\*/

proc sql noprint;

select count(distinct compress(name)) into :Ceiling

from &Disease.\_Current;

quit;

/\*Compress the names of sites for do loop processing\*/

data &Disease.\_Historical;

set &Disease.\_Historical;

name=compress(name);

run;

data &Disease.\_Current;

set &Disease.\_Current;

name=compress(name);

run;

/\*Create Subsets of the data for each site - Both a historical table and a current season table\*/

%do i=**1** %to &Ceiling;

data &&Site&i.**.\_**Historical;

set &Disease.\_Historical;

if upcase(name) ne upcase("&&Site&i") then delete;

run;

data &&Site&i.**.\_**Current;

set &Disease.\_Current;

if upcase(name) ne upcase("&&Site&i") then delete;

run;

%end;

\*---------------------------------------------------------------------------;

/\* MACRO STEP 4 - USE MACRO PROGRAMMING TO RUN THE WVAL ANALYSIS AT THE SITE LEVEL\*/

/\*RSV (for us) doesn't have enough historical data to calculate representative STD's and 10th percentiles so we use current data\*/

%if &Disease=RSV %then %do;

%do i=**1** %to &Ceiling;

proc univariate data=&&Site&i.**.\_**Current noprint;

var log\_virus;

output out=&&Site&i.**.\_**sum\_stats std=std p10=tenthperc;

run;

proc sql noprint;

select std, tenthperc

into :SiteSTD&i. , :SiteTen&i.

from &&Site&i.**.\_**sum\_stats;

quit;

%end;

%end;

/\*For our other pathogens we have enough data to derive insights from a historical table so see the example of this below. Notice how we pull the summary stats from the historical table now.\*/

%else %do;

%do i=**1** %to &Ceiling;

proc univariate data=&&Site&i.**.\_**Historical noprint;

var log\_virus;

output out=&&Site&i.**.\_**sum\_stats std=std p10=tenthperc;

run;

/\*Store those summary statistics in a macro to call to\*/

proc sql noprint;

select std, tenthperc

into :SiteSTD&i. , :SiteTen&i.

from &&Site&i.**.\_**sum\_stats;

quit;

%end;

%end;

/\*Create the variables required for the analysis and perform the wval calculations. This is the part where we refer to the macro variables we have been storing in memory.\*/

%do i=**1** %to &Ceiling;

data &&Site&i.**.\_**Current;

length level\_cat $12.;

set &&Site&i.**.\_**Current;

Site\_Baseline= &&SiteTen&i.;

Site\_STD\_DEV= &&SiteSTD&i.;

Site\_Deviation\_Log=(log\_virus-Site\_baseline)/Site\_STD\_DEV;

Site\_Deviation\_Linear=exp(Site\_deviation\_log);

if virus\_l=**.** then level\_cat='No Data';

else if Site\_deviation\_linear =< **1.5** then level\_cat='Minimal';

else if **1.5** < Site\_deviation\_linear =< **3** then level\_cat='Low';

else if **3** < Site\_deviation\_linear =< **6** then level\_cat='Moderate';

else if **6** < Site\_deviation\_linear =< **11** then level\_cat='High';

else if **11** < Site\_deviation\_linear then level\_cat='Very High';

run;

%end;

/\*Merge each of the seperate sites data into one table\*/

%do i=**1** %to &Ceiling;

proc append base=&Disease.\_WVAL\_Current data=&&Site&i.**.\_**Current force;

run;

%end;

proc sort data=&Disease.\_WVAL\_Current;

by sample\_date;

run;

\*---------------------------------------------------------------------------;

/\*MACRO STEP 5 - LOAD IN THE MMWRWEEK CALCULATOR CODE\*/

/\* Disclaimer: this segment of code below was something I found on StackOverflow (a website for asking coding questions/sharing helpful functions publicly)

/\* create a data set describing your intervals. Here, the year begins on the Sunday on or preceeding Jan 4 And the intervals are Sunday-Saturday \*/

data CDC\_Weeks(keep=begin end MMWRWEEK year);

start\_year = **2010**;

end\_year = **2025**;

do year = start\_year to end\_year;

/\* Sunday preceeding January 4th \*/

begin\_year = INTNX("WEEK",MDY(**1**,**4**,year),**0**);

next\_Jan\_4 = MDY(**1**,**4**,year+**1**);

do MMWRWEEK = **1** to **53**;

begin = INTNX("WEEK",begin\_year,MMWRWEEK-**1**);

end = INTNX("WEEK",begin,**0**,'E');

if (end LT next\_Jan\_4) then output;

end;

end;

/\* use a format to indicate that these are SAS DATE values \*/

format begin end MMDDYY10.;

run;

\*---------------------------------------------------------------------------;

/\*MACRO STEP 6 - JOIN THE MMWR WEEK VALUES TO THE WVAL TABLE\*/

/\* JOIN CDC MMWR WEEK CALCULATOR CODE WE CREATED ABOVE TO WVAL Current data TO BRING IN THE MMWR WEEK AND MMWR WEEK END DATA\*/

proc sql;

create table &Disease.\_WVAL\_Current1 as

select a.\* , b.MMWRWEEK, b.end as MMWR\_Week\_End

from &Disease.\_WVAL\_Current as a

left join Work.Cdc\_weeks as b

on a.Sample\_Date between b.begin and b.end;

quit;

\*---------------------------------------------------------------------------;

/\*MACRO STEP 7 - CREATE A TABLE FOR STATEWIDE VIRAL LEVELS IN THE WASTEWATER\*/

proc sql;

create table &Disease.\_CURRENT1 as

select a.\* , b.MMWRWEEK, b.end as MMWR\_Week\_End

from &Disease.\_CURRENT as a

left join Work.Cdc\_weeks as b

on a.Sample\_Date between b.begin and b.end;

quit;

/\*Add in a Current Statewide Virus/L concentration – Here I create both a median statewide value and a mean statewide value\*/

proc sql;

create table &Disease.\_Current2 as

select \*, median(virus\_L) as median\_virus\_L,

log(calculated median\_virus\_L) as log\_median\_virus\_l, sum(virus\_L)/count(name) as Mean\_State\_VirusperL,

log(calculated Mean\_State\_VirusperL) as log\_Mean\_State\_VirusperL

from &Disease.\_CURRENT1

group by mmwr\_week\_end;

quit;

/\*Limit to one observation per week\*/

proc sql;

create table &Disease.\_Current3 as

select distinct mmwrweek, median\_virus\_l, log\_median\_virus\_l, Mean\_State\_VirusperL, log\_Mean\_State\_VirusperL, MMWR\_Week\_End

from &Disease.\_Current2;

quit;

/\*Add in the value "Statewide" for name so we can merge it into the previously created dataset with the sites\*/

data &Disease.\_Current3;

set &Disease.\_Current3;

Name="Statewide";

Virus\_L=median\_virus\_l;

log\_Virus=log\_median\_virus\_l;

run;

\*---------------------------------------------------------------------------;

/\*MACRO STEP 8 - CREATE A TABLE FOR HISTORICAL STATEWIDE VIRAL LEVELS IN THE WASTEWATER – THIS STEP MAY LOOK LIKE DUPLICATED CODE FROM ABOVE, BUT NOTICE IT IS USING THE HISTORICAL DATA NOW.\*/

/\* JOIN CDC MMWR WEEK CALCULATOR CODE TO Historical data TO BRING IN THE MMWR WEEK AND MMWR WEEK END DATA\*/

proc sql;

create table &Disease.\_historical1 as

select a.\* , b.MMWRWEEK, b.end as MMWR\_Week\_End

from &Disease.\_historical as a

left join Work.Cdc\_weeks as b

on a.Sample\_Date between b.begin and b.end;

quit;

/\*Add in a Historical Statewide Virus/L concentration\*/

proc sql;

create table &Disease.\_historical2 as

select \*, median(virus\_L) as hist\_median\_virus\_L,

log(calculated hist\_median\_virus\_L) as log\_hist\_median\_virus\_L, sum(virus\_L)/count(name) as Hist\_Mean\_State\_VirusperL,

log(calculated Hist\_Mean\_State\_VirusperL) as log\_Hist\_Mean\_State\_VirusperL

from &Disease.\_historical1

where virus\_l ne **.**

group by mmwrweek;

quit;

/\*Limit to one observation per week\*/

proc sql;

create table &Disease.\_historical3 as

select distinct mmwrweek, hist\_median\_virus\_L, log\_hist\_median\_virus\_L, Hist\_Mean\_State\_VirusperL, log\_Hist\_Mean\_State\_VirusperL, MMWR\_Week\_End

from &Disease.\_historical2;

quit;

/\*Add in the value "Statewide" for name so we can merge it into the big dataset with the sites\*/

data &Disease.\_historical3;

set &Disease.\_historical3;

Name="Statewide";

run;

\*---------------------------------------------------------------------------;

/\* MACRO STEP 9 - USE MACRO PROGRAMMING TO RUN THE WVAL ANALYSIS AT THE STATEWIDE LEVEL\*/

/\*Calculate the Standard Deviation and 10th percentile for the state\*/

proc univariate data=&Disease.\_historical3 noprint;

var log\_Hist\_Mean\_State\_VirusperL;

output out=&Disease.\_historical\_sum\_stats std=std p10=tenthperc;

run;

/\*Store those summary statistics in a macro to call to\*/

proc sql noprint;

select std, tenthperc

into :StateSTD , :StateTen

from &Disease.\_historical\_sum\_stats;

quit;

/\*Create the variables required for the analysis and perform the wval calculations – this is similar to what we did with the site level data, but for our state activity determination we compare the current week median to the 10th percentile of the historical weekly averages / the std of the historical averages.\*/

data &Disease.\_Current3;

length level\_cat $12.;

set &Disease.\_Current3;

State\_Baseline= &StateTen;

State\_STD\_DEV= &StateSTD;

State\_Deviation\_Log=(log\_median\_virus\_l-State\_Baseline)/State\_STD\_DEV;

State\_Deviation\_Linear=exp(State\_Deviation\_Log);

if State\_Deviation\_Linear =< **1.5** then level\_cat='Minimal';

if **1.5** < State\_Deviation\_Linear =< **3** then level\_cat='Low';

if **3** < State\_Deviation\_Linear =< **6** then level\_cat='Moderate';

if **6** < State\_Deviation\_Linear =< **11** then level\_cat='High';

if **11** < State\_Deviation\_Linear then level\_cat='Very High';

run;

/\*Join the historical state avg columns to the current table\*/

proc sql;

create table &Disease.\_State\_data as

select a.\*, b.hist\_median\_virus\_L, b.log\_hist\_median\_virus\_L, b.Hist\_Mean\_State\_VirusperL, b.log\_Hist\_Mean\_State\_VirusperL

from &Disease.\_current3 a

left join

&Disease.\_historical3 b

on a.mmwrweek=b.mmwrweek;

quit;

proc sort data=&Disease.\_State\_data noduprecs;

by mmwrweek;

run;

/\*Append the Statewide Data into the WVAL Data\*/

proc append base=&Disease.\_wval\_current1 data=&Disease.\_current3 force;

run;

proc sort data=&Disease.\_wval\_current1;

by mmwr\_week\_end;

run;

/\*Here I do some county level aggregations, but we don’t really use this at the moment so you can probably omit this step if you wanted to.\*/

proc sql;

create table &Disease.\_wval\_current2 as

select \*, avg(virus\_l) as County\_Virus,

avg(log\_virus) as County\_Log\_Virus,

avg(site\_deviation\_linear) as County\_Deviation\_Linear

from &Disease.\_wval\_current1

group by mmwr\_week\_end, County;

quit;

/\*This is where we create the final table we will export to excel. It contains the Activity level for the state and the activity for the sites. The state value is compared to historical state data on the weekly level, and the site data is compared to it’s own historical data to determine its individual level.\*/

proc sql;

create table &Disease.\_wval\_current3 as

select a.\*, b.median\_virus\_l, log\_median\_virus\_l,

b.hist\_median\_virus\_L, b.log\_hist\_median\_virus\_L, b.Hist\_Mean\_State\_VirusperL, b.log\_Hist\_Mean\_State\_VirusperL, b.State\_Baseline, b.State\_STD\_DEV, b.State\_Deviation\_Log, b.State\_Deviation\_Linear

from &Disease.\_wval\_current2 a

left join

&Disease.\_state\_data b

on a.mmwrweek=b.mmwrweek and a.mmwr\_week\_end = b.mmwr\_week\_end;

quit;

proc sort data=&Disease.\_wval\_current3 noduprec;

by MMWR\_Week\_End mmwrweek name;

run;

\*---------------------------------------------------------------------------;

/\*The first sort here gets rid of any duplicate values, the second lists the data by ascending calendar week\*/

proc sort data = &Disease.\_WVAL\_Current3 noduprecs;

by ID Sample\_Date;

run;

proc sort data = &Disease.\_WVAL\_Current3;

by mmwr\_week\_end;

run;

\*---------------------------------------------------------------------------;

/\* MACRO STEP 11 - DATA CLEANING TO GET RID OF UNECESSARY DATASETS\*/

/\*Get rid of uneccessary datasets. This just cleans up your work library view so you only see the datasets you need to, and not all the ones you created for temporary use previously.\*/

/\*Delete all the site tables\*/

%do i=**1** %to &Ceiling;

proc datasets nolist lib=work;

delete &&Site&i.**.\_**Current &&Site&i.**.\_**Historical &&Site&i.**.\_**sum\_stats;

quit;

%end;

proc datasets nolist lib=work;

delete &Disease.\_current &Disease.\_Historical All\_Regions All\_Regions1 &Disease.\_WVAL\_Current CDC\_Weeks &Disease.\_statewide &Disease.\_historical1 &Disease.\_wval\_current1 &Disease.\_wval\_current2 &Disease.\_state\_data &Disease.\_current1 &Disease.\_current2 &Disease.\_historical2 &Disease.\_historical3 &Disease.\_historical\_sum\_stats &Disease.\_current3 flowdata &Disease.\_Percent\_Change\_Stats Normalize &Disease.\_All\_Data;

quit;

\*---------------------------------------------------------------------------;

/\* MACRO STEP 12 - EXPORT DATA FOR VISUAL MODELING IN POWER BI\*/

/\*WVAL Data\*/

proc export data=&Disease.\_WVAL\_Current3 outfile="C:\Users\JohnXW\OneDrive - State of Oklahoma\Shared Documents - OSE\_IDPR\_IDSC Landing Page - EID\Wastewater\SAS Analysis\WVAL Code Excel Data\&Disease. Wval Data.xlsx"

dbms=xlsx Replace;

run;

\*---------------------------------------------------------------------------;

**%mend** Disease;

\*---------------------------------------------------------------------------;

/\*LIST OF MACROS TO RUN\*/

/\*Run This Macro for Covid Data\*/

%***Disease***(libname=Disease, ImportSheet=COVID\_Master, Disease=Covid)

/\*Run This Macro for Flu B Data\*/

%***Disease***(libname=Disease, ImportSheet=Influenza B\_Master, Disease=FluB)

/\*Run This Macro for Flu A Data\*/

%***Disease***(libname=Disease, ImportSheet=Influenza A\_Master, Disease=FluA)

/\*Run This Macro for RSV Data\*/

%***Disease***(libname=Disease, ImportSheet=RSV\_Master, Disease=RSV)

/\*Run This Macro for NoroVirus\*/

%***Disease***(libname=Disease, ImportSheet=Norovirus\_Master, Disease=Norovirus)