ArboMAP: Arbovirus Modeling and Prediction to Forecast Mosquito-Borne Disease Outbreaks

User Guide (v2.0)

Justin K. Davis and Michael C. Wimberly
(justinkdavis@ou.edu, mcwimberly@ou.edu)
Geography and Environmental Sustainability, University of Oklahoma

Updated November 22, 2019

Contents

What you need to know about ArboMAP

ArboMAP is a program used to model and predict cases of vector-borne disease. Here we consider human cases of the mosquito-born West Nile virus, but ArboMAP will likely work with any data set in which:

- there are multiple (>3) years of infection data from years in which the pathogen can be considered endemic, after any introductory years in which the pathogen is probably rapidly exploiting naive populations and should not yet be considered established
- the disease has distinct transmission and quiescent seasons, including an initial, annual exponential growth phase when the pathogen is beginning to spread after a period of relatively few cases. For the moment, ArboMAP assumes that the transmission season occurs during the calendar year and essentally ceases November through March. In settings where a transmission season crosses the December-January boundary, the code will require modification.
- there is reason to believe that incidence responds to measurable environmental indices, almost certainly including temperature and probably some measure of moisture in the environment (e.g. precipitation or humidity)
- cases of disease are assigned to districts (states, counties, etc.) and cases are not too rare - every modeled district should have at least one case over the period of study, as the model will assume a district is permanently immune if no cases have been observed there
- some measure of pathogen in the environment is available; e.g. here we use the rate at which pools of mosquitoes test positive for the virus.

In what follows we will occasionally use the language of our West Nile virus studies (mosquitoes, counties, etc.) but the modeling approach does not rely on those details and any appropriate data set can be substituted.

The ArboMAP user guide consists of four main sections. The first section describes how to install the necessary software needed to run the system. The second section describes the steps that need to be taken every year before the beginning of the WNV season to prepare for forecasting. The third section description the steps that need to be taken each time a new forecast is generated. The final section describes the outputs produces by ArboMAP and explains how to interpret the forecast charts and maps.

What you need to know about WNV and modeling

West Nile virus (WNV) in South Dakota is our example in this document. WNV circulates primarily among mosquitoes and birds, but occasionally spills over into human or other hosts. In humans, the majority (~80%) of infections show no symptoms and are not diagnosed except perhaps accidentally during blood donation. Some small proportion of individuals will develop symptoms such as headache, fever, and rash, but in about 1 out of every 100 infected individuals, the disease becomes neuroinvasive and can cause debilitating illness and death.

WNV was first introduced to the US in 1999 on the east coast, after which it spread in a westward-moving wave across the US. The early years produced thousands of cases as WNV invaded new regions. Birds were immunologically naive, so the virus decimated entire avian populations and spilled over into humans in an outbreak that lasted several years in some states.

In South Dakota, there were more than a thousand diagnosed cases in 2002-2003, which implies that many more were actually infected, since most infected individuals do not show symptoms and are never diagnosed. After a number of years of active transmission, case numbers declined and it looked like the disease might vanish from circulation after 2011, in which there were only two cases reported.

However, in 2012 there was another outbreak, with hundreds of cases in South Dakota. Because of these massive fluctuations in annual case counts, there is a need to predict the magnitude and locations of WNV in advance to support proactive mosquito control and disease prevention activities. Thus, a key scientific question is whether there are there any environmental or entomological indicators that could tell us, in advance, how bad a year is likely to be.

In fact, yes. Work by Dr. Michael Wimberly's research group has shown that a relatively simple statistical model, relying on weather and mosquito infection data, can be used to predict the risk of infection on a district-week basis. For more background, you will want to consult relevant journal articles, which are listed at the end of this document. In particular, the paper by Davis et al. (2018 in Acta Tropica) provides details about the underlying statistical model.

Setting up for the initial run

A note on prepackaged data

ArboMAP comes packaged with example human, mosquito, and weather data. The weather data are real, but the human and mosquito data were generated randomly according to model estimates from our WNV study in South Dakota. Hence, the data "look like" real human and mosquito data, but are synthetic and *should not be used* for any actual scientific study. They are included merely to make sure that all software is installed and settings are correct; you will probably want to run the system first with these synthetic data before trying your own data.

Some accounts you'll need

You will need some measure of pathogen in the environment; obtaining these data is usually a substantial undertaking requiring multiple locations and participants, so this will probably be stored in some central database requiring a login. Here, we use a website to which participants around the state added their mosquito infection data.

You will need some measure of the environment so that relationships between disease and environment can be modeled. Here we use the gridMET dataset, downloaded from Google Earth Engine (GEE). GEE is a cloud-based platform for processing satellite remote sensing images and other geospatial datasets. Instead of downloading the data and processing it yourself, GEE uses cloud-based computers and algorithms to do in half an hour what might take a standard desktop computer a week to calculate.

To access GEE, you will first need a gmail account at **mail.google.com**, so that you can access Google Drive documents at **drive.google.com**. Next, visit GEE at **earthengine.google.com** and click on "Sign Up" to request an account. You receive confirmation at your gmail address, and any downloads of weather data will be stored in the Google Drive of the account.

Finally, you will need outcome data, usually human cases. These data are probably best obtained by data-sharing agreements with a health department. Up-to-date case counts are not necessary - ArboMAP models the current year based on all previous years, and does not rely on case counts in the current year, which are in any case usually misleadingly low. The location of residence (or transmission) and the date of onset of symptoms (or diagnosis or detection) are required. The date of onset is rounded to week in this analysis and the location should be a larger district (e.g. county) rather than a specific address.

Some software you'll need

R

R (**www.r-project.org**) is a statistical programming language that runs all of our analyses and produces reports and documentation, including this document. It is free and has a wide variety of packages built by users all around the world to do essentially any statistical analysis you can imagine.

It is easiest to download R from CRAN (**cloud.r-project.org**). Click on the link for "Download R for Windows." Choose "base," then "Download R for Windows." Run this file and install R on your system. Use the default settings for the installer.

RStudio

RStudio (**www.rstudio.com**) is a user-friendly GUI for the statistical programming language R that greatly simplifies a number of tasks for the programmer. Navigate to the site, click on Download, and choose the RStudio Desktop (Open Source License) - that is, the free version. Run the appropriate installer, very likely the Windows Vista/7/8/10 installer.

Run RStudio and close it at least once after you've installed MiKTeX as described in the next section. RStudio may not be able to find MiKTeX unless it's had a chance to search.

MiKTeX

MiKTeX (**www.miktex.org**) is an implementation of LaTeX, which is what allows us to automatically produce PDF reports at the end of all the calculations. Navigate to the site, click on Download -> Windows and then download and run the installer. Default options will suffice, but make one change: "Install missing packages on-the-fly" should be changed to "yes." You will initially install an incomplete copy of MiKTeX, and it will be updated automatically whenever you run the code. This means that the first time you run the code, it might take quite a while.

An alternative is to choose the Net Installer from the "All Downloads" page, and to choose "Complete Installation." This will take a great deal of time, and you should only do it on a fast connection, but it means you will only ever have to do this once. Expect that this will take a few hours - if you can afford the time to do this, it's worth getting all the downloads done. You will probably be asked to select a source/server/mirror from which to download your files - you can choose the 0-Cloud, which will automatically find a fast download site for you, or you can manually pick somewhere close to you physically.

Setting up the directory structure

ArboMAP requires a variety of files to run, and the following directory structure is the easiest way to organize everything that's necessary. If you download the repository and compile the main ArboMAP code in RStudio with Ctrl+Shift+K, you will draw from data sources already in place and will not need to modify the contents of these directories.

Files and settings

The ArboMAP User's Guide.Rmd file, which you will open in RStudio, contains a preface (shown below) with settings that tells the program where to find the various pieces of data used to make predictions. Errors in these filenames are a common reason that the code will refuse to run.

- `graphicoutputdir` tells the program where to store its graphical outputs. You may want to use some of the graphics in the PDF for your own purposes, so they will be stored in PNG format in this directory.
- `humandatafile` contains the human WNV cases
- `maxobservedhumandate` tells the program which is the last human case that should be used in modeling. If you happen to know that there are some cases in 2018 and you're predicting 2018, then you should set this to the end of 2017. We do not use the current year's data in making predictions, and trying to do so will mess up the predictions. This is an extra safeguard in case you do accidentally update the case file to include data from this year.
- `weekinquestion` tells the program which week you'd like to obtain predictions for. Weeks in this program begin on Sunday and run until the next Saturday, and the weekinquestion date will automatically be rounded to the previous Sunday. For example, if 2018-07-16 happens to be a Monday, then the week examined is 2018-07-15 (Sunday) through 2018-07-21 (Saturday).
- `weatherpathstr` indicates the directory where all the weather data CSV files are stored

- `weathersummaryfile` is the name of the file that the program will create to summarize all the CSV files. Since it's possible you've downloaded a year many times (especially the current year), the program has to create a summary file and stores it with the rest of the weather data. Other than this extra file and all the CSV files, there should be nothing else in this directory.
- `mosqfile` contains the mosquito testing data
- `stratafile` tells us where to find the file containing the data to split up the state into four regions to model the mosquito infection rate (see below)
- `districtshapefile` points to the SHP file for mapping districts
- `compyear1` and `compyear2` indicate which two years in the past we'd like to compare the current year to in one of the graphics.
- `var1name` and `var2name` indicate which two weather variables you'd like to use as predictors in the model. By default, these are set to mean daily temperature from the gridMET set (tmeanc) and the vapor pressure deficit (vpd).

`Maxobservedhumandate` should be updated at the beginning of every season to point to the end of last year, so that only human cases in previous years are used to condition the model's predictions. `Weekinquestion` should be updated weekly.

```
# where do we want the outputs?
graphicoutputdir <- ".\\graphical outputs\\"
fullcasematoutputdir <- ".\\case matrix with estimates\\"
mosqmatoutputdir <- ".\\mosquito matrix with estimates\\"

# where are the human data located?
humandatafile <- ".\\human case data\\simulated human case data.csv"

# what is the date of the last human case we're willing to believe?
# probably, cut this off at the end of last year
# DO NOT use any human cases from the year you're modeling
maxobservedhumandate <- as.Date("2017-12-31", "%Y-%m-%d")

# which week are we producing the graphs for?
weekinquestion <- as.Date("2018-08-15", "%Y-%m-%d")

# where are the weather csv files stored?
weatherpathstr <- ".\\weather data\\"
# what is the name of the summary file to be created?
weathersummaryfile <- "weather data summary file.csv"

# which variables do you want to use?
var1name <- "tmeanc"
var2name <- "vpd"

# where are the mosquito test files located?
mosqfile <- ".\\mosquito data\\simulated mosquito tests.csv"

# which district stratification scheme are we using?
stratafile <- ".\\strata\\17-04-20 - classified strata - classic.csv"

# where is the districtshapefile
districtshapefile <- ".\\shapefile\\cb_2014_us_county_5m - in EPSG 5070 - only SD.shp"

# to which two other years do we want to compare the current year's predictions?
compyear1 <- 2012
compyear2 <- 2017
```

Choosing your models

The `modelformulas` object contains a list of formulas used to model the data. Currently, the naming convention describes whether cubic (cub) or thin-plate (tp) splines are used. Then, the distributed lags are either fixed (fx) or seasonally-varying (sv). Then, we either use anomalized (anom) or non-anomalized (nonanom) environmental data.

So, for example, the "cub-sv-nonanom" model is the model we used for our most recent publications. The "tp-sv-anom" model takes advantage of improvements in the mgcv library to use thin-plate splines to model distributed lags which vary over the season, over anomalized weather data. The tp-sv-anom model currently has the best fit out of all the listed models on our historical data, but this may not be the case for all users.

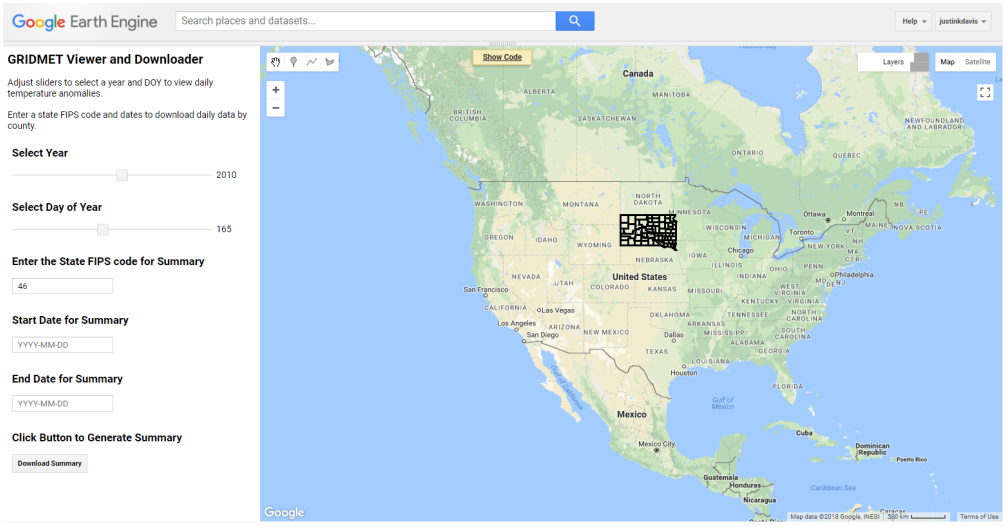
Typically for new users, we will recommend the simple "cub-fx-nonanom" model, which is most numerically stable and unlikely to fail, and the "tp-sv-anom" model for what is likely to be the best fit, but requires more computational power and may fail if the data do not tell a clear story.

The user has the option to run all listed models simultaneously by writing `modelnames <- names(modelformulas)` instead of `modelnames <- c("cub-sv-nonanom", "tp-sv-anom")` or some other list of models. This is not advisable unless the user is on a decently-powerful computer with a number of cores available for processing. We use the parallel processing of the mgcv::bam to fit models quickly, and ArboMAP should not be run on a busy computer or a laptop on battery power. Running the two default models takes approximately five minutes on an older desktop, and is much faster on a server with many available cores. If the run-time becomes excessive, choosing simpler models should help, but running individual models in separate instances of ArboMAP is unlikely to produce any benefit.

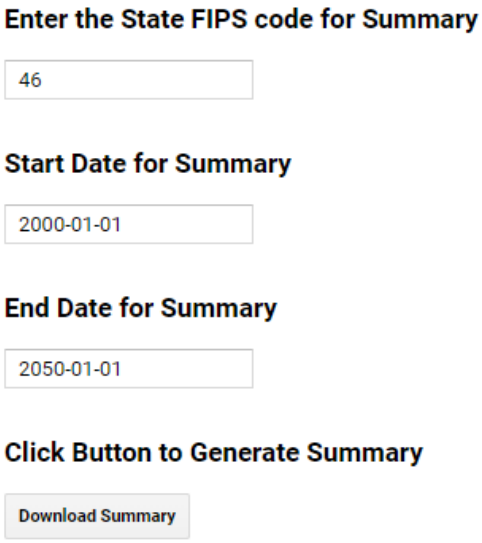
Obtaining data at the beginning of the season

Obtaining the gridMET data

ArboMAP requires weather data to make its predictions. You will use likely Google Earth Engine (GEE) to download historical weather data beginning a full year before your human case data (e.g. if you begin modeling in 2001, you will need to download all data beginning in 2000). If you want to download all available historical data, it will not hurt the calculations; simply begin in 1950). You will only need to download historical data once; every other time you run the program, you will only need to download the most recent year's weather data from GEE. Navigate to code.earthengine.google.com and sign up for an account if you have not already.

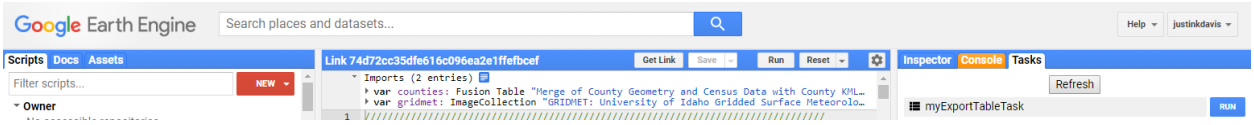


Click on the textured bar above "Show Code" and drag down to view the code. Paste into the code window the contents of the GRIDMET_downloader_v1_0.js file, included in the ArboMAP repository. Then click on "run" to render the GUI. You don't need to modify any of the code, but this will allow you to see the Tasks tab, which will allow you to download the data. Set the "Start Date for Summary" to YYYY-01-01, where YYYY is the year of data you wish to begin in, and set the "End Date for Summary" to some point in the future. You can set the End Date to 2050-01-01, and you will be given the most up-to-date weather data.

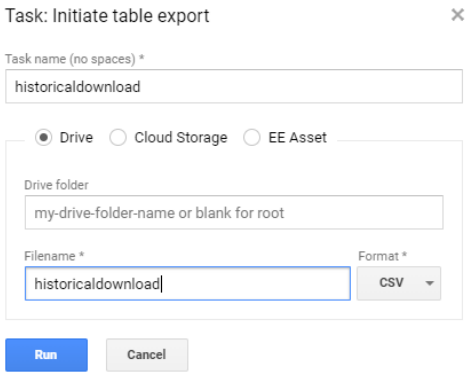


You can change some of the other options, if you'd like to modify your download. If you're not using the data for South Dakota, for example, which has FIPS code 46, you can choose another state from the list of codes (**click here for codes**). You will need to use the two-digit code (e.g. Colorado is 08, not 8).

Click on "Download Summary." This will create a task called myExportTable Task under the Tasks tab. Click on Run. If you do see this or any of the code, click on the textured bar directly above the map and drag down. This should reveal the code windows.



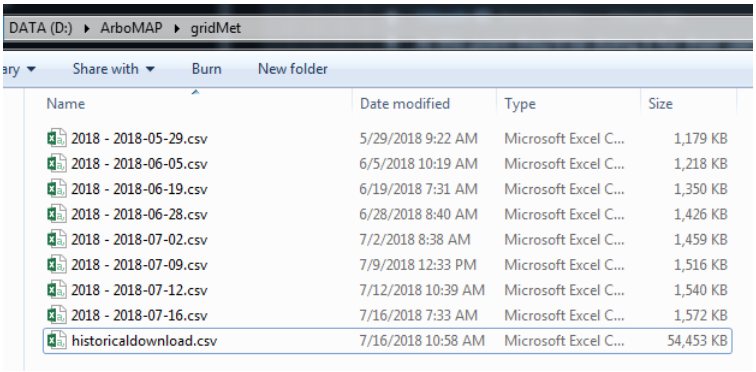
Name the task "historicaldownload" and name the file "historicaldownload" - note that there are no spaces. Click on "Run" in this window in the Tasks tab. If the Tasks tab is not visible, remember to pull down the bar directly above the map.



Processing can take half an hour for this initial data pull, depending on who else is using the system, and very rarely your job can be cancelled - you might receive an error about "too many objects," in which case you should try again later or split the download into individual years. Once it's done, there will be an extra file in your Google account's drive at **drive.google.com**.

historicaldownload.csv	me	12:58 PM me	53 MB
2018 - 2018-07-16.csv	me	9:33 AM me	2 MB
2018 - 2018-07-12.csv	me	Jul 12, 2018 me	2 MB
2018 - 2018-07-09.csv	me	Jul 9, 2018 me	1 MB
2018 - 2018-07-02.csv	me	Jul 2, 2018 me	1 MB
2018 - 2018-06-28.csv	me	Jun 28, 2018 me	1 MB
2018 - 2018-06-19.csv	me	Jun 19, 2018 me	1 MB
2018 - 2018-06-05.csv	me	Jun 5, 2018 me	1 MB
2018 - 2018-05-29.csv	me	May 29, 2018 me	1 MB

Download this file (right-click -> download) and save it in a directory with all the other weather data; by default this is "D:/ArboMAP/weather data/". You will eventually add updates for weather data in the current year (2018 at the time of writing), as discussed below, and these are the files beginning with "2018 - ". These updates are stored as extra CSV files and, along with the historical download, give you a complete picture of the daily weather in every district.



Note that it doesn't matter how many times you download a date's weather data, as long as it appears at least once in the data. Every single one of these 2018 files contains weather data for 2018-01-01, for example. Additional measurements will not affect the predictions, and any data that are represented more than once are simplified before inclusion into the model.

If you're not using GEE

Although GEE is recommended for a variety of reasons, you might obtain your environmental data from elsewhere. Below is format expected of any CSV file that you might instead supply. District is a human-readable name identifying the area where the weather measurement was taken. The "doy" indicates the day of the year (1-366). The remaining columns are measurements of environmental conditions. If your measurements are not daily (e.g. 7-day precipitation totals), they will need to be resampled to that temporal resolution.

district	doy	year	tminc	tmeanc	tmaxc	pr	rmean	vpd
Grant	1	2009	-16.1498	-8.89499	-1.64013	0	80.65256	0.131497
Roberts	1	2009	-16.4839	-9.20541	-1.92694	0	78.83587	0.137992
Marshall	1	2009	-15.7183	-9.05239	-2.38651	0	81.24488	0.134333
Day	1	2009	-15.3654	-9.05889	-2.7524	0	79.93273	0.12634
Deuel	1	2009	-15.2488	-8.41722	-1.58563	0	81.47671	0.125203
Codington	1	2009	-14.9043	-8.67153	-2.43874	0	79.7313	0.122761

Obtaining the human case data

You will only need to update the human case data file once at the beginning of each season. For example, to make WNV forecasts for 2018, the model is calibrated using human case data through the end of 2017, and 2018 case data are not incorporated until data are finalized after the end of the 2018 WNV season. Only two pieces of information are needed for every human case: a date and a district. The date is the symptom onset date (in the case of clinical cases) or the date of blood donation (in the case of viremic blood donors). The district is the patient's district (e.g. county) of residence. Each row represents a single case.

The CSV file you use should that contains the human case data should have a unique column named "date" that is formatted MM/DD/YYYY, and a column named "district" that contains standard names for location of residence (or transmission, if available). The names of districts are simplified by the code, so that any of the following are equivalent: BROOKINGS, Brookings, brookings, Brookings County, BROOKINGS COUNTY. These will all be reduced to "brookings" in the output.

It does not matter which other columns are present, as long as each row contains at least these two values. We have randomly generated some human case data that resemble the trends observed in WNV in SD. The following picture shows the expected format.

	A	B
	county	creationdate
2	brown	7/4/2004
3	brown	7/11/2004
4	brown	8/8/2004
5	yankton	8/29/2004
6	brown	6/26/2005
7	kingsbury	7/10/2005
8	lake	7/10/2005
9	brown	7/17/2005
10	brookings	7/24/2005
11	turner	7/24/2005
12	brown	7/31/2005
13	kingsbury	7/31/2005

Save this file in its own directory. By default this is "D:/ArboMAP/human/simulated human case data.csv".

Do not include any human case data from the year you are modeling. If you're making predictions for 2018, your model should be based on all human case data up to December 2017. If you include any human case data for 2018 - for example, if you know that there were a handful of cases already in mid-July 2018 and update the human case data file during the season - then the model will assume that human case data for 2018 are complete. This is almost certainly incorrect, since some cases are only reported weeks or months after diagnosis, and human case counts during the season will therefore underestimate the actual disease burden.

Formatting the vector infection data

Here we use a list of tested mosquito pools to quantify the amount of pathogen present in the environment. All tests should be present in a single CSV file, with the following columns in the following formats. The pool_size and species columns can be replaced by 1 and "default" respectively, but should be present. The wnv_result column can be 0 for a negative test or 1 for a positive test. Save this file in its own directory. By default this is "D:/ArboMAP/mosquito data/simulated mosquito tests.csv".

county	col_date	wnv_resul	pool_size	species
codington	6/3/2004	0	1	Culex tarsalis
codington	6/4/2004	0	1	Culex tarsalis
codington	6/5/2004	0	1	Culex tarsalis
codington	6/7/2004	0	4	Culex tarsalis
codington	6/11/2004	0	1	Culex tarsalis
codington	6/12/2004	0	2	Culex tarsalis
codington	6/16/2004	0	22	Culex tarsalis
pennington	6/16/2004	0	3	Culex tarsalis
pennington	6/16/2004	0	1	Culex tarsalis
codington	6/17/2004	0	3	Culex tarsalis
codington	6/18/2004	0	13	Culex tarsalis
codington	6/20/2004	0	8	Culex tarsalis
codington	6/22/2004	0	5	Culex tarsalis
pennington	6/22/2004	0	1	Culex tarsalis
codington	6/24/2004	0	36	Culex tarsalis
codington	6/24/2004	0	17	Culex tarsalis
codington	6/25/2004	0	2	Culex tarsalis
codington	6/26/2004	0	2	Culex tarsalis
codington	6/28/2004	1	5	Culex tarsalis

A note on standardized place names

ArboMAP was designed to model WNV on a district-week basis, but can work with any geographical unit as long as each area is assigned a unique, stable identifier. Ideally, you will work with a standard set of names (or even a numerical code like the FIPS). District names might not be not completely standardized in your sources; e.g. some sources will list "Brookings County" while others will simply use "Brookings".

ArboMAP reduces all names by deleting "County" and "Parish", removing all spaces from the district name, and putting the name in lowercase. Therefore, "BROOKINGS COUNTY" becomes "brookings" and "Charles Mix County" becomes "charlesmix". This gives the greatest chance that all places will be recognized in all data sources. However, incorrect spellings and invalid date formats (e.g. DD/MM/YYYY instead of MM/DD/YYYY) cannot be fixed by the program and will result in errors.

You will need to be sure that the shapefile you use contains a “NAME” field with human-readable names.

The typical run during the season

Updating the data during the season

Before you run the model during the WNV season, you will want to be sure you have the most up-to-date weather and mosquito data. These updates are simpler than the updates that need to be run once per year at the beginning of the WNV season, but will need to be performed every time you want to make an updated prediction.

Updating the gridMET weather data

Follow exactly the process described above, but only download data for the current year. For example, if you are modeling in 2018, set the start and end dates to 2018-01-01 and 2019-01-01. This will include any gridMET data available for 2018. Name the task "2018" and name the file "2018 - 2018-07-16" for example, if

you downloaded the data on 2018-07-16. Save this file in the gridMet directory with the historicaldownload.csv file and the other updates to 2018. Eventually, you will have a large number of these update files along with the original "historicaldownload.csv" file that contained all available past data.

Updating the human case data

Do not update the human case data during the season. If you learn that there was a case last week, for example, this should not yet be included as a new line in the "human case data.csv" file. Next year, when the historical human case data are updated, this new human case will be included in the file, but during a year the "human case data.csv" file should remain untouched.

Updating the mosquito data

Use the process described above to log in to the SDMIS website. This time, however, only download the active view, and replace "active_testing.csv" with the new file. During the season, the "archive_testing.csv" file will not change, since the historical data should not be revised.

Setting up the Rmd file

You will be given a file called "ArboMAP.Rmd", which should be placed in "D:/ArboMAP/". This is an Rmarkdown file, which does all of the calculations of modeling and produces a PDF in the end to summarize the outputs. Open RStudio and open this file.

Running the code

Open the Rmd file in RStudio, change the weekinquestion, and press ctrl+shift+K. This will tell RStudio to compile the PDF. It should not take very long - on a slow laptop, modeling takes around 10 minutes at most. If it takes substantially longer than this, it is likely that there is an error somewhere in the data. At the end, you will see a PDF in the same directory as the Rmd file, and a number of graphics will be stored in the directory you requested.

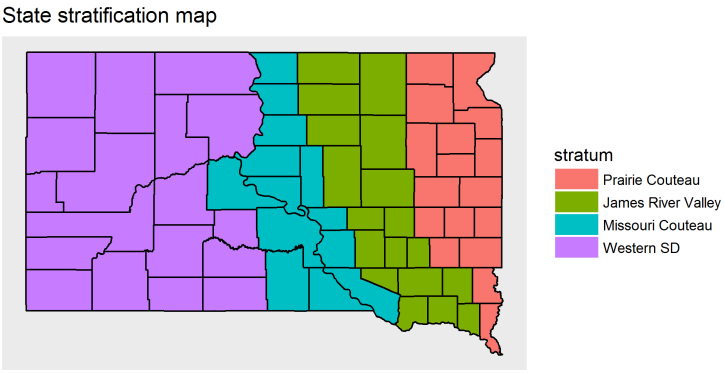
Interpreting the output

Weather data

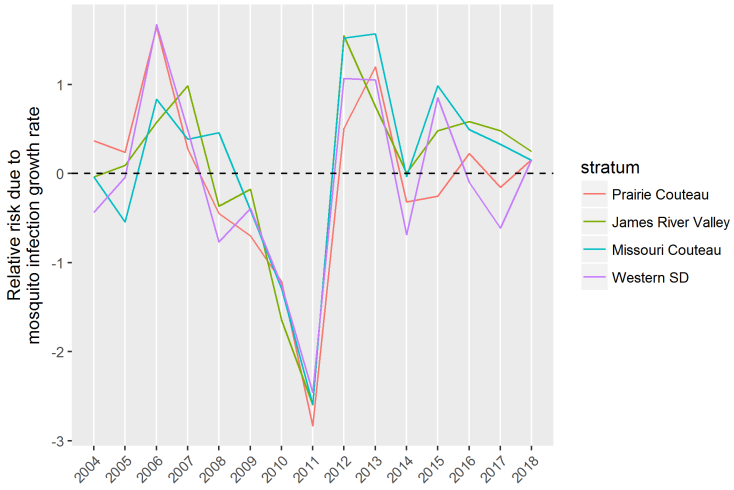
Each forecasting run will generate a report containing a number of graphics to interpret. The raw and anomalized environmental data for all years will be shown, including measurements during the current year. The anomalized data are data from which the daily average has been subtracted, so the anomaly is the measurement above or below what is normal for that point in the year.

Mosquito data

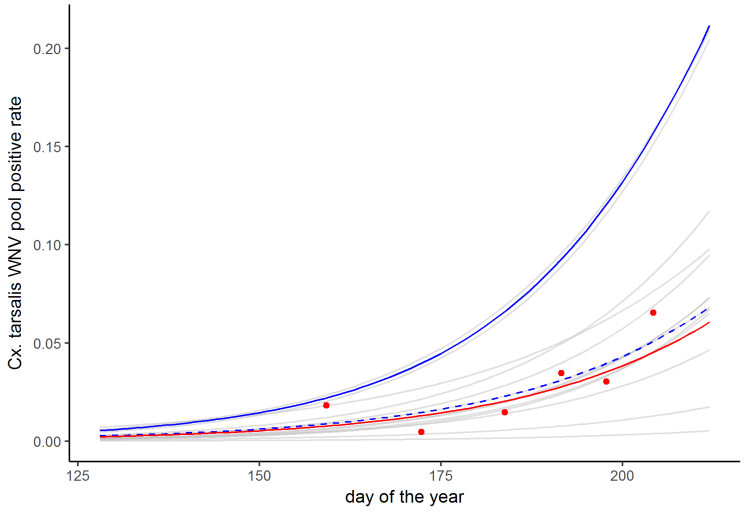
The model of WNV also summarizes the mosquito infection growth rate (MIGR) for multiple strata within the state. Every winter, the virus goes into hiding. In the early season, WNV begins replicating and spreading among birds and mosquitoes. The MIGR is a measure of how quickly that's occuring. We cannot estimate the MIGR for every district, but instead split the state into four strata, shown below, and estimate the MIGR within the strata. If the map does not resemble this, especially if the map is entirely grey, it's likely that the program cannot locate the strata file.



Below we show the estimates of the MIGR per year, per stratum. If the line is above 0, then mosquito infections are growing in pools more quickly than in the average year, and more human WNV should be expected. If lower, then risk to humans is lower. Note that the state does fluctuate up and down as a whole (e.g. in 2011 when barely any positive mosquito pools were found, and only two human cases were diagnosed), but sometimes the strata do differ (e.g. the MIGR in western SD was relatively low in 2014, but was relatively high in 2015).



Note that the MIGR is a single number, derived from a model that fits all of a year's mosquito infection data. Below, we show observations from 2018 (red dots), along with the curve that best fits these data (red), and samples from two other years, 2012 (blue) and 2017 (blue, dashed). The faster this line rises, the higher the MIGR, and the higher the risk.



Censoring chart

It will typically be the case that there are long stretches of time in which there are no human infections. Trying to model these makes most algorithms unstable - if you try to model too many 0s, you are likely to see odd results in your model fit statistics, AIC, etc. Hence, we only model cases within a certain time period - in the default, synthetic data, this is between weeks 24 and 42 of the year. This range can be tightened or relaxed by adjusting the `humancasealpha` parameter in the settings, which retains 100(1-alpha)% of the cases to model.

Multi-year WNV forecasting chart

Next are the model outputs. For each week of a year, the model estimates what proportion of districts will report at least one human case based on fits to historical data (black). Each model specified in the list at the beginning of the code will be assigned one line.

Current-year WNV forecasting chart

Next are model outputs for the year you're estimating and the two years you've selected for comparison. The solid, vertical line indicates the week of the year for which predictions are produced.

Positives-to-cases chart

ArboMAP models positive district-weeks, which are either 0 if there were no cases in that district in that week, or 1 if there was at least one case. We can attempt to convert this to total case counts, by fitting a smooth curve (red, dotted) to the relationship between positive district-weeks and total case counts in a year (each dot). Predictions are summarized in the table below.

Current-week WNV absolute risk map

Next are the absolute probabilities of each district reporting at least one human case during the week in question. If the district is darkest blue, then the model says there is no probability of reporting a case during the week you've selected. This is often the case in the early or late seasons, especially in the lower-population districts. If a district is brightest red, then the model says the district will definitely report at least one human case this week. Each model will have a map.

Current-week WNV relative risk map

Next are the relative probabilities. A district in the early season may have low risk because the virus is not yet circulating, but is that lower or higher than usual, for that district, during that week of the year? Brookings

County on May 15th 2018 will have low risk, but that risk may be higher than usual when compared to Brookings on May 15th in 2004-2017? If a district is red, this means that risk is predicted to be substantially higher than average in that district, during the week in question. If blue, then lower. If yellow, then average. It is important to remember that red and blue do not mean high or low absolute risk. They mean high*er* or low*er than usual*. A district that is lower than usual can still have many cases in a week; e.g. Brown County in early August can be lower than usual, but will still likely report cases.

The predictions of all models are averaged to produce this map.

Estimated dependence functions

Because ArboMAP allows you to run a variety of models with different forms, the output in this series of graphs may vary wildly depending on your chosen formulas. Additionally, some of the outputs may be difficult to interpret. In all cases, above each graph you will see "model: cub-sv-nonanom" or some other model name which indicates which model the dependence function comes from. Then:

- s(doy) indicates a smooth over the day of the year (doy). This is usually a cyclical term used in anomalized models, which models the general trend over all years in the sample. This is relative - the lower the function at some day of the year, the less likely you are to see human WNV cases on that day of the year, all else being equal. Generally, this should range between -5 and 5 - if it goes beyond this range, it is likely that you are modeling too many empty district-weeks and should raise your humancasealpha so that more empty weeks are cut off.
- s(lag):variable is a single distributed distributed lag function, showing the dependence of risk now (lag=0) on environmental data some time in the past. Var1/2 are whichever variables you have chosen (temperature and vapor pressure deficit) in the settings. As an example, if the estimated distributed lag is positive at lag=120, then the environmental variable four months ago correlates with an increase in human cases today. If you used anomalized environmental data, these variables will be prefixed with "anom_".
- te(lag, doymat):variable indicate how the risk today (lag=0) depends on environmental data at some point in the past. Var1/2 are whichever variables you have chosen (temperature and vapor pressure deficit) in the settings. If you used anomalized environmental data, this will have the "anom_" prefix. If you have chosen a seasonally-varying model, then these will depend on the day of the year and three different dependence functions will be shown. For example, precipitation in the winter (near doy=0) will not have much of an effect on human risk, whether two days (lag=2) or three months (lag=90) ago. If these lines are essentially all the same, you might achieve a better fit with fixed (fx) rather than seasonally-varying (sv) distributed lags.

Selecting covariates to use

Previously, ArboMAP included another Rmd file to facilitate selection of covariates. Unfortunately, as the number of options increases, it is no longer feasible to run all combinations of model structure and environmental covariates. We suggest you run at least two models - the models given in the default settings are reliable and test with a variety of environmental covariates. Typically, you will want to begin with one temperature and one moisture index - mean daily temperature (tmeanc) and the vapor pressure deficit (vpd) have served us well, and are not a bad place to start.

Troubleshooting

Filenames and unrecognized escapes

The most common error will be an error in filenames. If you are told that a file is not found, it is likely that you have not correctly told the Rmd file where to look. Another common error with filenames concerns the slashes: if you see something like the following, then it is because you have used a single \, which is not

acceptable, instead of \\ or / in your filename. Note that we have written “d:\work” rather than “d:\\work” or "d:/work".

```
D:/work/git/WNV/WNV/WNV_modeling_-_main.Rmd
Line 38 Error: '\w' is an unrecognized escape in character string starting ""D:\w" Execution halted
```

```
# where are the human data located?
humandatafile <- "D:\work\\git\\WNV\\packaged WNV\\human\\17-11-16 - reconciled human WNV.csv"
```

No estimates, blank maps

Unfortunately, if the Rmd file refuses to produce estimates (e.g. the red line is missing in the graphs or the maps are completely blank) then it’s likely that a file is missing, is in an improper format, or there are extra files where they do not belong; e.g. only gridMet files should go in the gridMet data directory.

Bad estimates

This is usually because of bad mosquito infection data. You will want to carefully check the mosquito infection data to ensure they are in the correct format. You will often be able to see this on the graph of the mosquito infection growth rate - a year in the past will dip down to very low levels of infection, but this does not seem historically accurate. Another problem can occur if you enter district names incorrectly. If the data tell the program that a case occurs in a district that does not exist, then this case will not be counted, and a district’s risk will be artificially low. If this happens with enough districts (e.g. if every district is followed by a space in the data file) then a whole year’s risk will be 0.

Relevant scientific papers

- Davis, J. K., G. P. Vincent, M. B. Hildreth, L. Kightlinger, and M. C. Wimberly. 2018. Improving the prediction of arbovirus outbreaks: a comparison of climate-driven models for West Nile virus in an endemic region of the United States. Acta Tropica 185: 242-250.
- Davis J. K., Vincent G. P., Hildreth M. B., Kightlinger L., Carlson C., and M. C. Wimberly. 2017. Integrating Environmental Monitoring and Mosquito Surveillance to Predict Vector-borne Disease: Prospective Forecasts of a West Nile Virus Outbreak. PLoS Currents Outbreaks. 2017 May 23. Edition 1. doi: 10.1371/currents.outbreaks.90e80717c4e67e1a830f17feeaaf85de.
- Wimberly, M. C., A. Lamsal, P. Giacomo, and T. Chuang. 2014. Regional variation of climatic influences on West Nile virus outbreaks in the United States. American Journal of Tropical Medicine and Hygiene 91: 677-684.
- Wimberly, M. C., P. Giacomo, L. Kightlinger, and M. B. Hildreth. 2013. Spatio-temporal epidemiology of human West Nile virus disease in South Dakota. International Journal of Environmental Research and Public Health 10: 5584-5602.
- Chuang, T., C. W. Hockett, L. Kightlinger, and M. C. Wimberly. 2012. Landscape-level spatial patterns of West Nile virus risk in the northern Great Plains. American Journal of Tropical Medicine and Hygiene 86: 724-731.
- Chuang, T., G. M Henebry, J. S. Kimball, D.L. VanRoekel-Patton, M. B. Hildreth, and M. C. Wimberly. 2012. Satellite microwave remote sensing for environment modeling of mosquito population dynamics. Remote Sensing of Environment 125: 147-156.
- Chuang T., and M. C. Wimberly. 2012. Remote Sensing of Climatic Anomalies and West Nile Virus
- Chuang, T. M. B. Hildreth, M. B., D. L. VanRoekel, and M. C. Wimberly. 2011. Weather and land cover influences on mosquito populations in Sioux Falls, South Dakota. Journal of Medical Entomology 48: 669-679.
- Wey, C. L., J. Griesse, L. Kightlinger, and M. C. Wimberly. 2009. Geographic variability in geocoding success for West Nile virus cases in South Dakota, USA. Health & Place 15: 1108-1114.
- Wimberly, M. C., M. B. Hildreth, S. P. Boyte, E. Lindquist, and L. Kightlinger. 2008. Ecological niche of the 2003 West Nile virus epidemic in the northern Great Plains of the United States. PLoS One 3:

e3744.